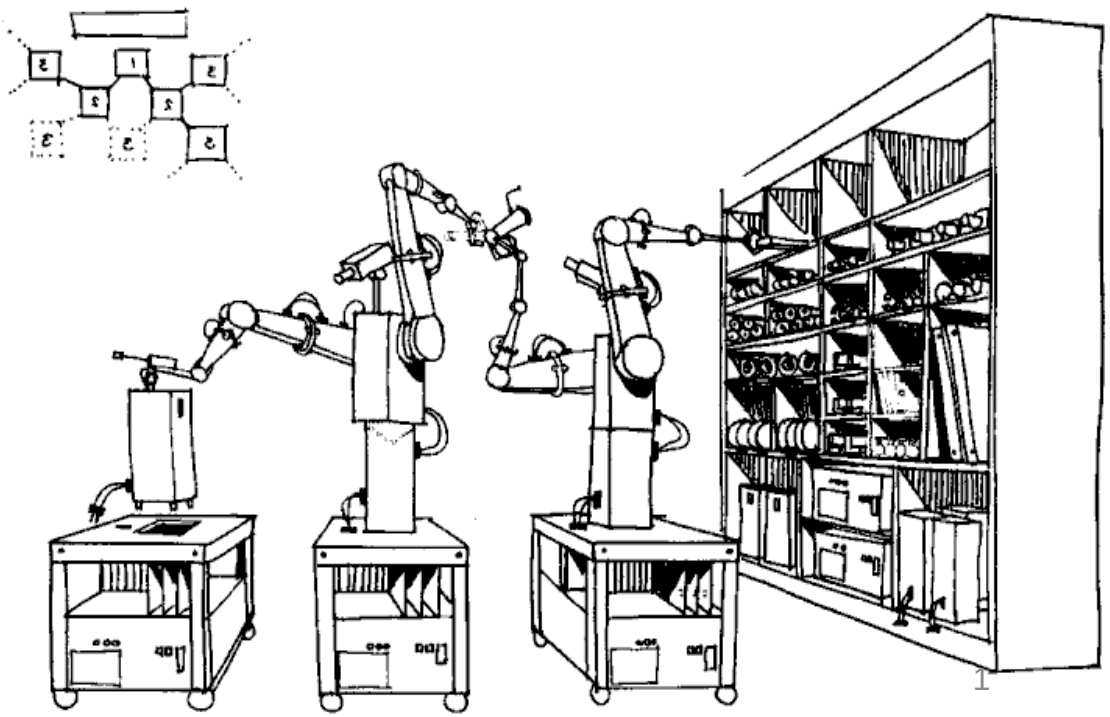
Introduction to data science and applications



What's possible with data and analytics?

Facebook can predict break-ups?

Facebook Can Predict With Scary Accuracy If Your Relationship Will Last

02/14/2014 10:37 am ET | Updated Feb 14, 2014

http://www.huffingtonpost.com/2014/02/14/facebook-relationship-study_n_4784291.html

Possible

Facebook can suggest interesting trending news stories?

FACEBOOK'S TRENDING FEED ATTRACTS FAKE STORIES AFTER PURGE OF HUMAN EDITORS

By **Saqib Shah** — August 30, 2016 10:42 PM

http://www.digitaltrends.com/social-media/facebook-trending-fake-news-blunders/

Facebook fires trending team, and algorithm without humans goes crazy

Module pushes out false story about Fox's Megyn Kelly, offensive Ann Coulter headline and a story link about a man masturbating with a McDonald's sandwich

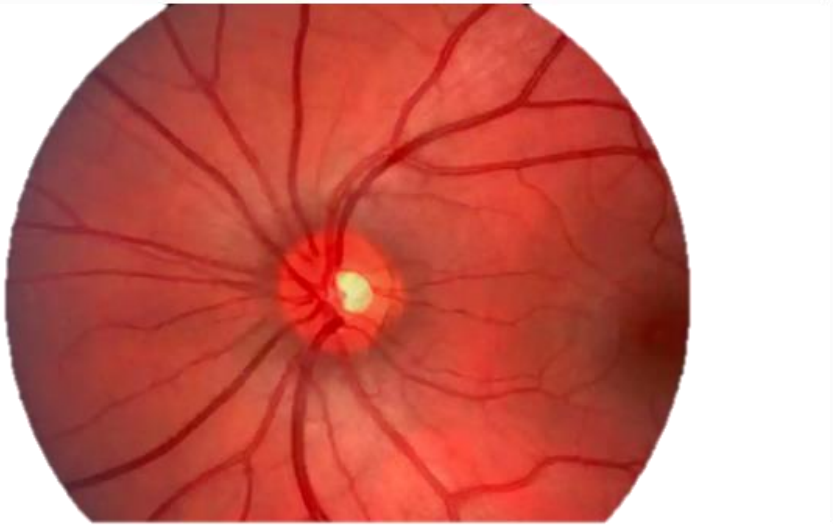
Impossible

(Not without some human help)

Google can detect early signs of blindness?

Google DeepMind pairs with NHS to use machine learning to fight blindness

'Deep learning' research company will use 1m anonymised eye scans to train a neural network to identify early signs of degenerative eye conditions



https://www.theguardian.com/technology/2016/jul/05/google-deepmind-nhs-machine-learning-blindness

Possible

(Perphaps… soon)

Google can detect flu epidemics?

DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

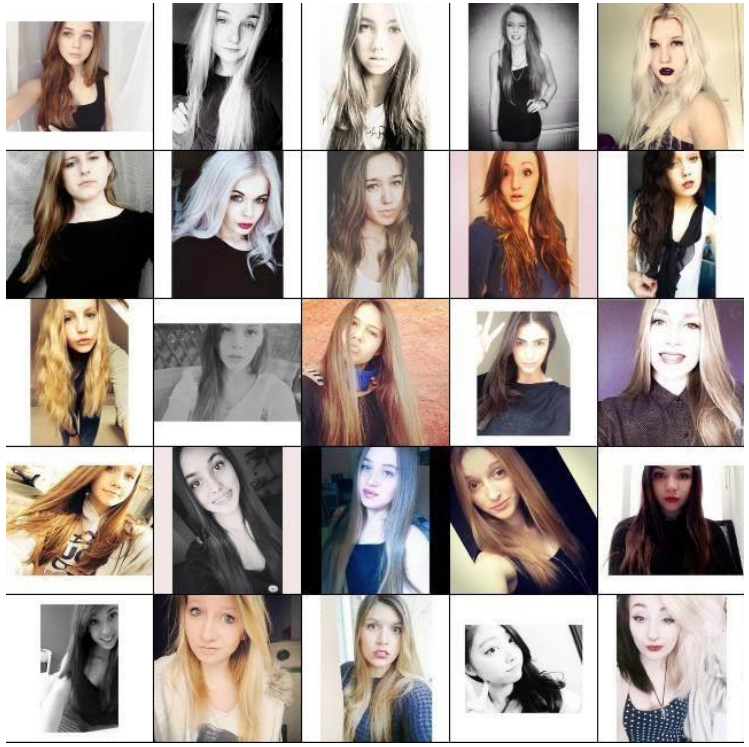
WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Impossible

(Not without strict maintenance)

An algorithm knows whether your selfie will succeed or fail?



http://karpathy.github.io/2015/10/25/selfie/

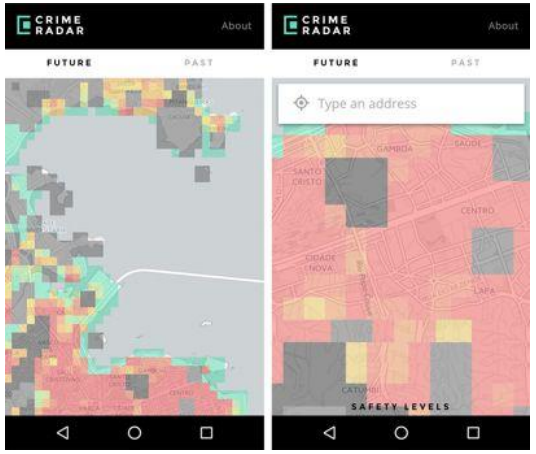
Possible

Data can predict crime before it happens?

CrimeRadar is using machine learning to predict crime in Rio

The software carves the city into sectors of 250 square metres and predicts crimes based on time and place

http://www.wired.co.uk/article/crimeradar-rio-app-predict-crime



'Minority Report'-style technology to predict crime in China

http://www.telegraph.co.uk/technology/2016/03/09/minority-report-style-technology-to-predict-crime-in-china/

Possible

(Working on it)

Facebook can predict your check-ins?

The goal of this competition is to predict which place a person would like to check in to.

https://www.kaggle.com/c/facebook-v-predicting-check-ins/

Possible

Satellite images can improve investments?

RISE OF "QUANTAMENTAL" INVESTMENT FUNDS

Leading hedge fund BlackRock, for example, is using satellite images of China taken every five minutes to better understand industrial activity and to give it an independent reading on reported data.

Traditionally, there have been two main types of actors in the financial world—traders (including high-frequency traders), who look to make money from massive volumes on many small transactions, and investors, who look to make money from a smaller number of larger bets over a longer time. Investors tend to care more about the underlying assets involved. In the case of company stocks, that usually means trying to understand the underlying or fundamental value of the company and future prospects based on its sales, costs, assets, and liabilities and so on.

Possible

It’s a big data world!

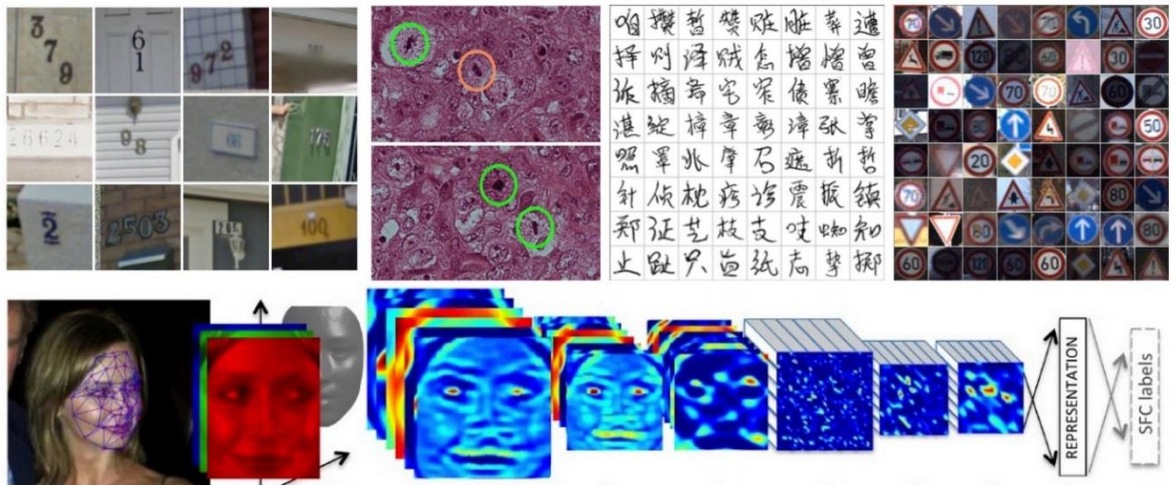
Banking

Insurance

Fraud

Marketing

Healthcare



Crime

Logistics

Real estate

Education

Retail

There’s no denying the fact

Data contains **value** and **knowledge**

But to extract it, you need to be able to

Store	**Manage**	**Analyze**
it	it	it

Now that you’re all hyped up

- Establish a **common vocabulary**
- Introduce **basic concepts**, techniques
- Highlight some **challenges**, key things to keep in mind

What is data science?

Data Science ≈ Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Knowledge Discovery (from Databases) ≈
Machine Learning ≈ Business Analytics ≈ Business Intelligence

Given some data, discover patterns or provide predictions that are:

- **Valid**: hold on new data with some certainty, i.e. **generalizable**
- **Useful**: should be possible to act on the item, i.e. **actionable**
- **Unexpected**: non-obvious to the system, i.e. **interesting**
- **Understandable**: humans should be able to interpret the pattern

Valid, generalizable?

https://neil.fraser.name/writing/tank/

Over time, seasonal effects, overfitting, sub-groups, regional differences…

Useful, actionable?

- “What if we built a super-smart artificial brain and no one cared?”
- What is we can detect credit card fraud with 100% accuracy… but it takes 30 minutes per check?
- Our model improves recommendations but is costly to keep running
- Our model predicts churn… but how to stop it?

Business question, implementation, maintenance costs, ease-of-use…

Unexpected, interesting?

- Customers who spend more churn less… duh!
- Frequent patterns or rather rare, decisive patterns?
- Optimize towards right-turns for UPS drivers
- Weird instead of big data

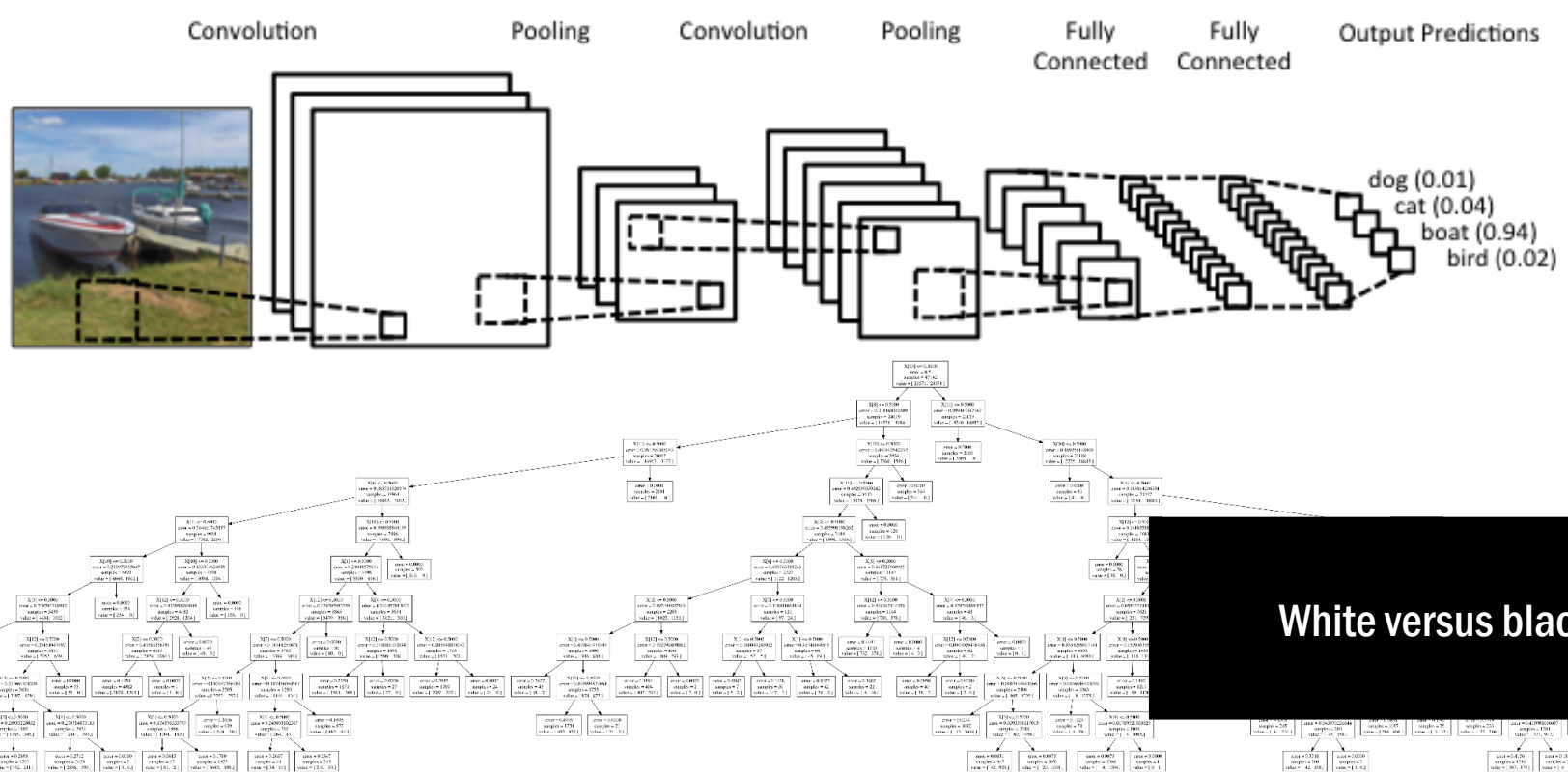
WEIRD SMALL DATA HAS ITS BENEFITS AND ITS RISKS

More than simply pop-economics, *Freakonomics* (2005) showed how unusual yet good-quality data sources can be valuable in creating insights. Assiduous record-keeping of the accounts of an honesty system cookie jar in an office place revealed that people stole most during certain holidays (perhaps due to increased financial and mental stress at these times); access to drug gangster bookkeeping accounts explained why many drug dealers live with their grandparents (they are too poor to move out); and massive public school records from Chicago showed parental attention to be a key factor in students' academic success.

https://www.fastcompany.com/3063110/the-rise-of-weird-data

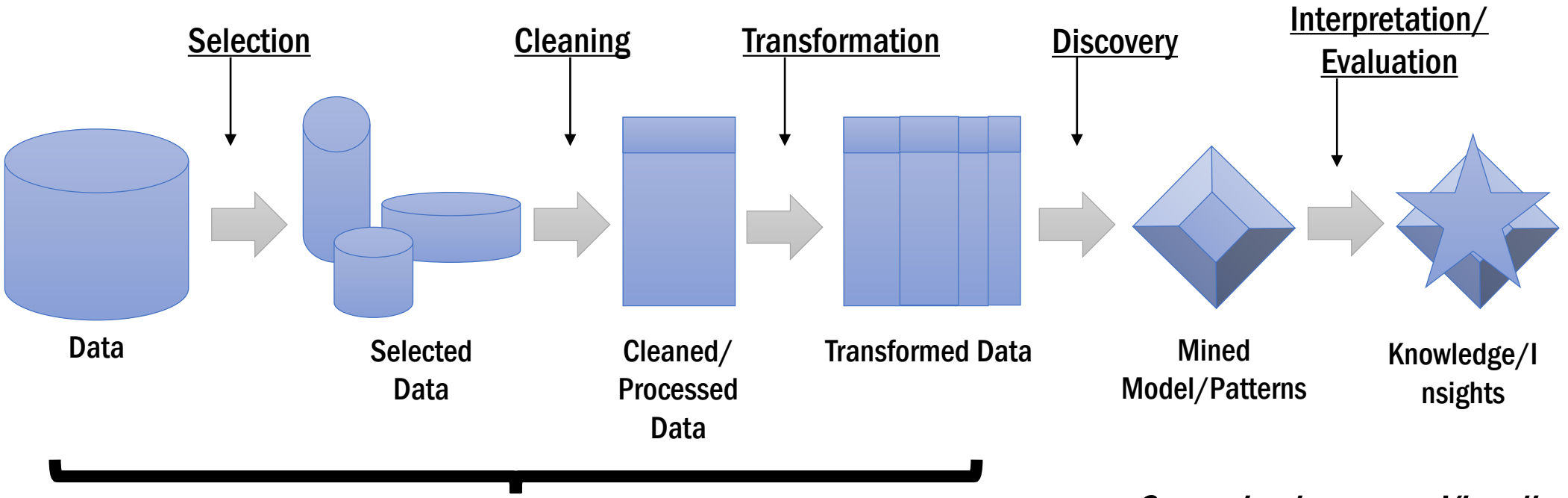
Not always a bad thing though, balance between trust and discovery…

Understandable?

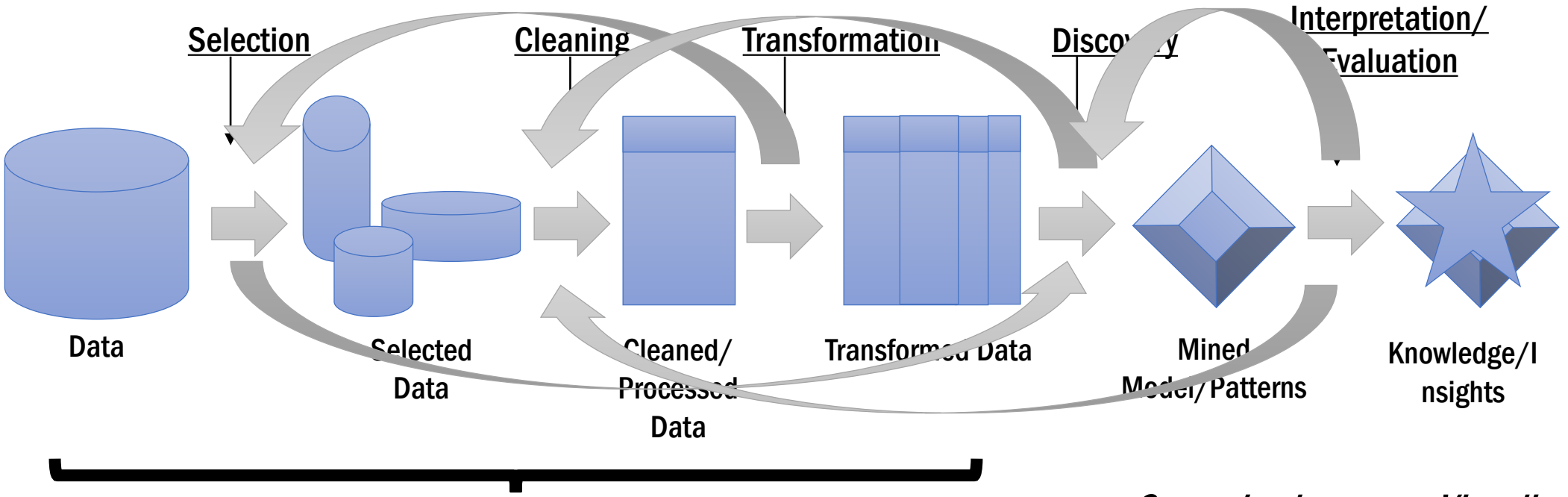


Black box vs. white box. Why? Related to trust, validity…

The data science process



The **real** data science process

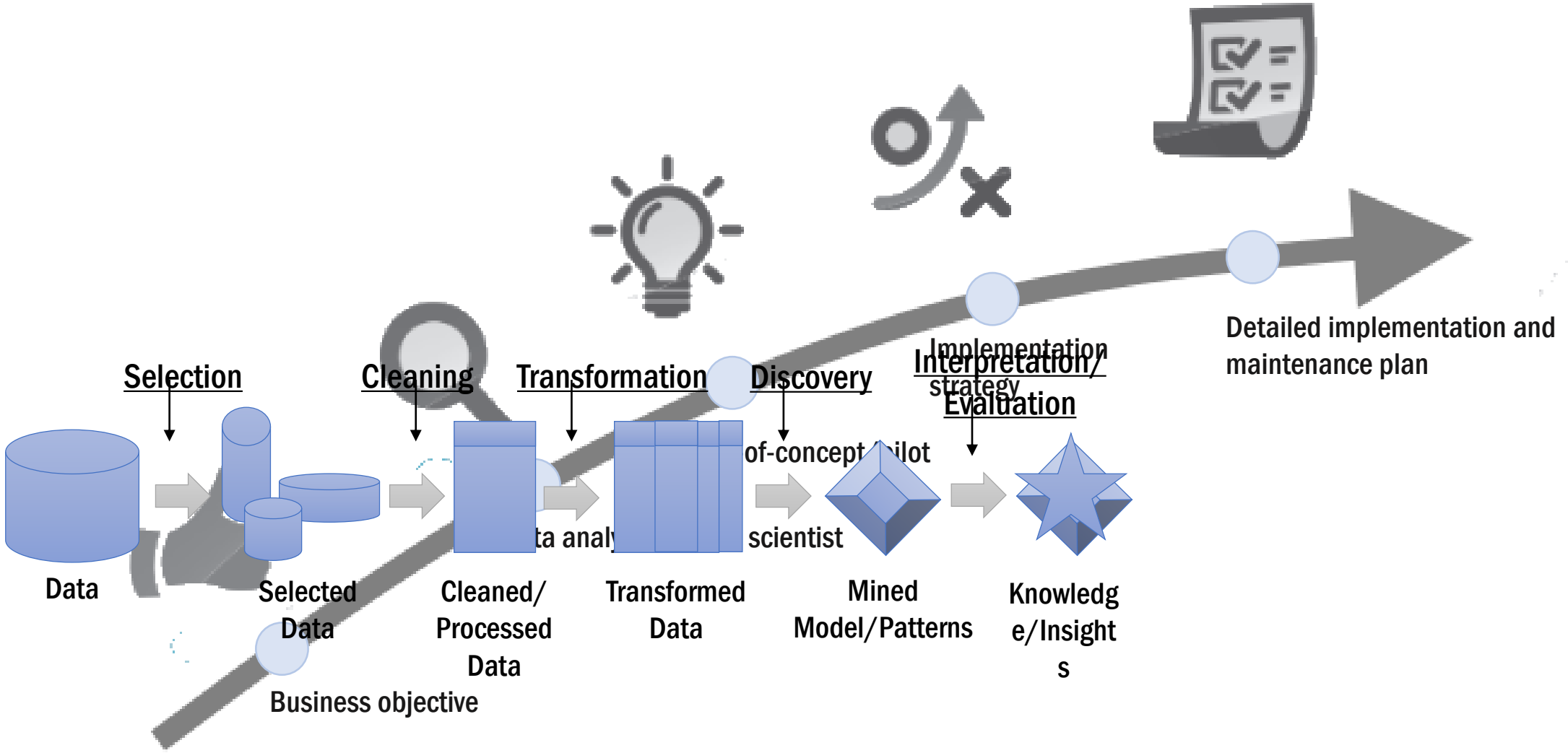


Preprocessing: sampling, de-noising, normalization, feature selection, dimension reduction, feature engineering

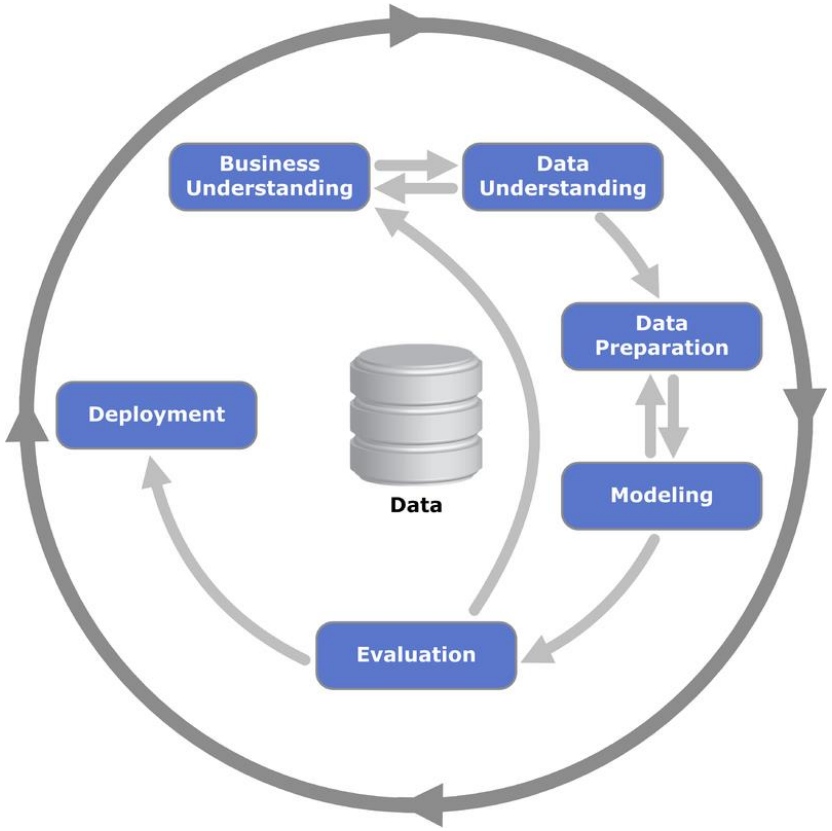
Supervised
Unsupervised

Visualisation
Validation

The **full** data science process



The data science process



CRISP-DM

SEMMA (SAS Institute)

S: Sample (Training, Validation, Test)

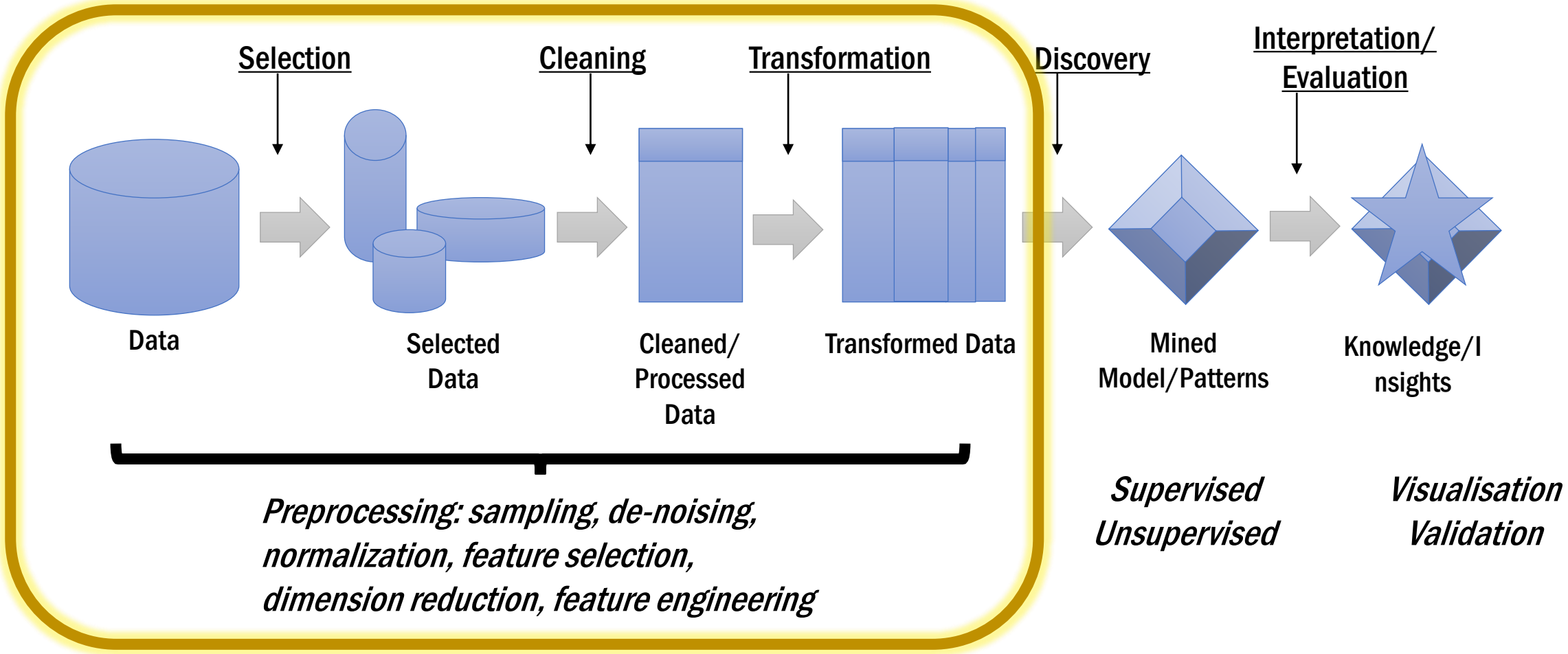
E: Explore (get an idea of the data at hand)

M: Modify (select, transform)

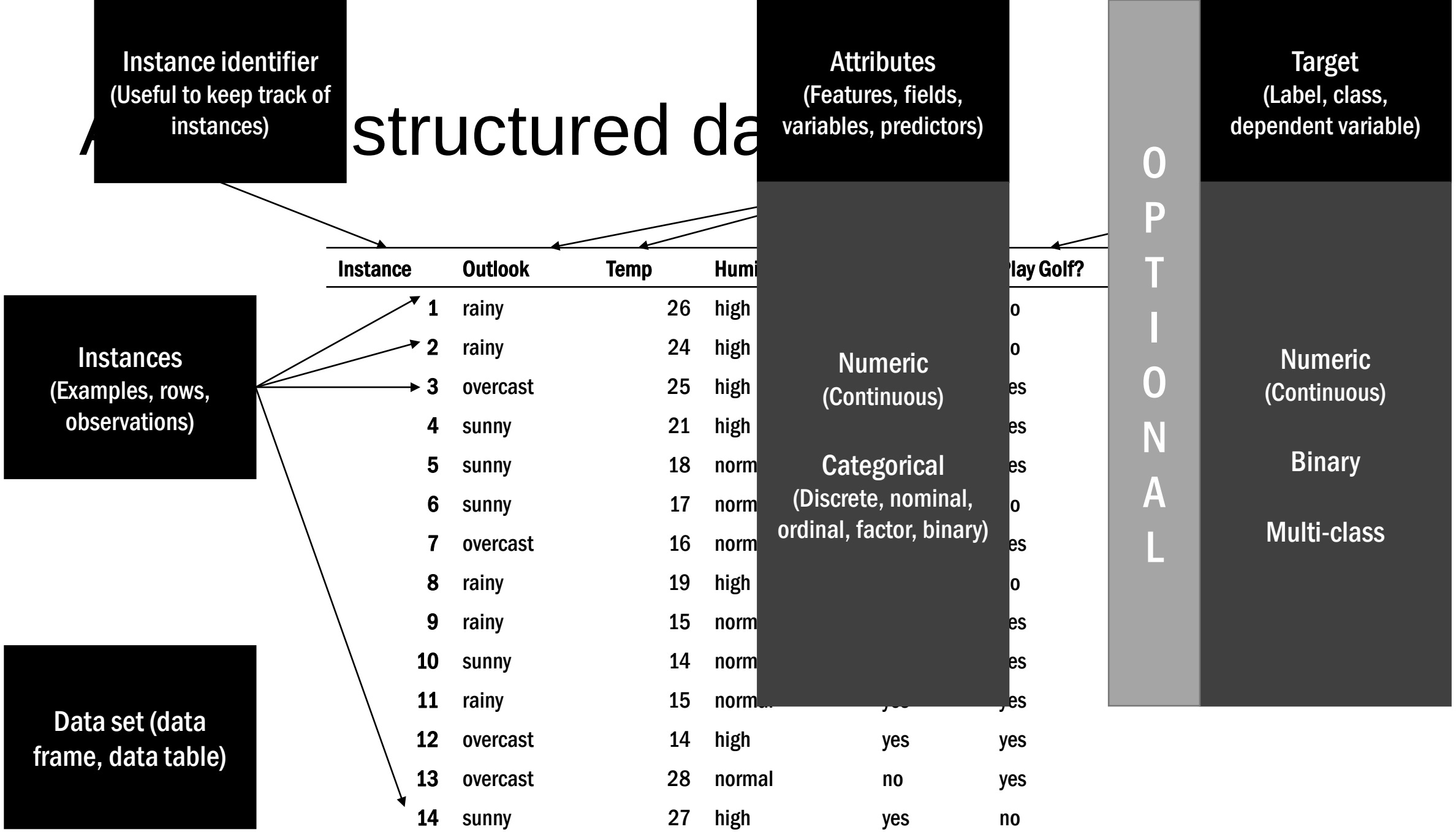
M: Model (create data mining model)

A: Assess (validate model)

The data science process



A structured da



Instance	Outlook	Temp	Humi		lay Golf?
1	rainy	26	high		o
2	rainy	24	high		o
3	overcast	25	high		es
4	sunny	21	high		es
5	sunny	18	norm		es
6	sunny	17	norm		o
7	overcast	16	norm		es
8	rainy	19	high		o
9	rainy	15	norm		es
10	sunny	14	norm		es
11	rainy	15	norma		es
12	overcast	14	high	yes	yes
13	overcast	28	normal	no	yes
14	sunny	27	high	yes	no

Preprocessing

Selection

Cleaning

Transformations

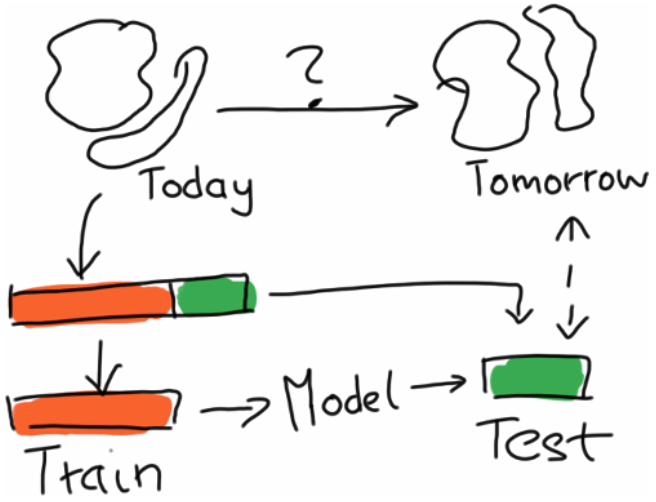
Feature selection

Feature extraction

Sampling

Preprocessing: selection

- Combining different data sources, data frames into one data set for use
- As data mining can only uncover patterns actually present in the data, target data set must be large enough to contain these patterns
- Setting aside a “hold-out” set: keep piece of data separate to verify your model works with unseen data
- Also: initial explorations and visualisation



Preprocessing: cleaning

- Detect errors/duplications in the data and rectify them if possible:

  Vanden Broucke, vandenbroucke, VANDEN/BROUCKE, …

- Data transformation: convert formats to common representation:

  Male, Man => M
  True, yes, ok => 1

Preprocessing: fixing missing values

Most techniques cannot deal with missing values out of the box

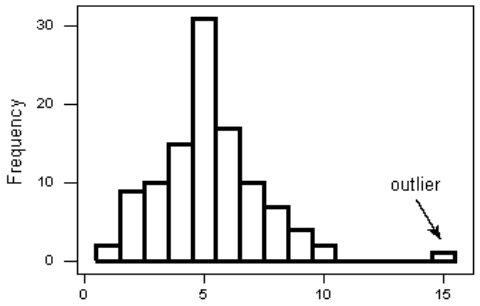
- Non applicable versus non disclosed
- **Delete row:** e.g. when too many missing values exist for that row
- **Replace (impute):** estimate missing value, replace e.g. by mean, median or mode
- **Keep:** include a separate missing indicator variable

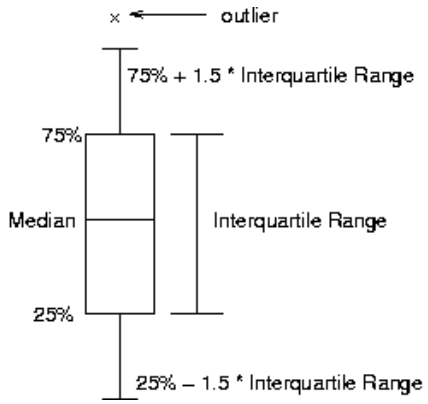
Instance	Outlook	Temp	Humidity	Windy	Play Golf?
1	rainy	26	high	no	no
2	rainy	24	high	yes	no
3		25	high	no	yes
4	sunny		high	no	yes
5	sunny			no	yes
6	sunny	17	normal		no
7	overcast	16	normal	yes	yes
8	rainy	19		no	no
9			normal	no	yes
10	sunny	14	normal		yes
11		15	normal	yes	yes
12	overcast	14	high	yes	yes
13	overcast	28		no	yes
14	sunny	27	high	yes	no

Preprocessing: fixing outliers and duplicates

Outliers are extreme observations

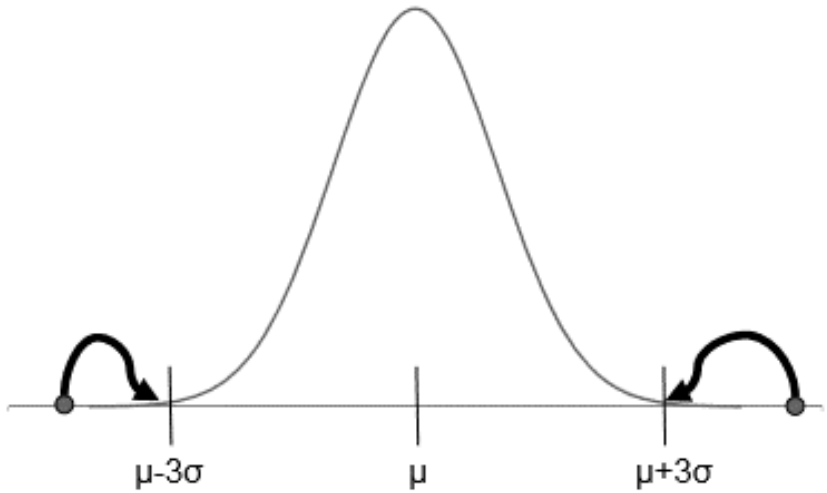
Age = 241
Income = 1 241 471 euro
Temperature = 51 degrees Celcius





- **Valid** observations: just very extreme outliers
- **Invalid** observations: due to errors, measuring mistakes, data entry error
- Difference between outlier **detection** and **treatment** (consider as missing, clip, create additional feature)

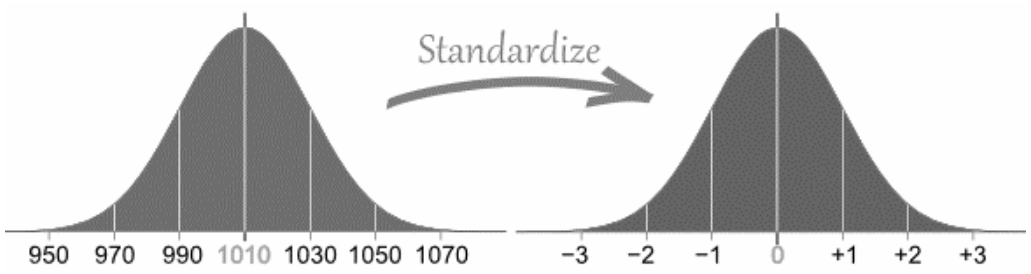
Duplicates: also valid or invalid

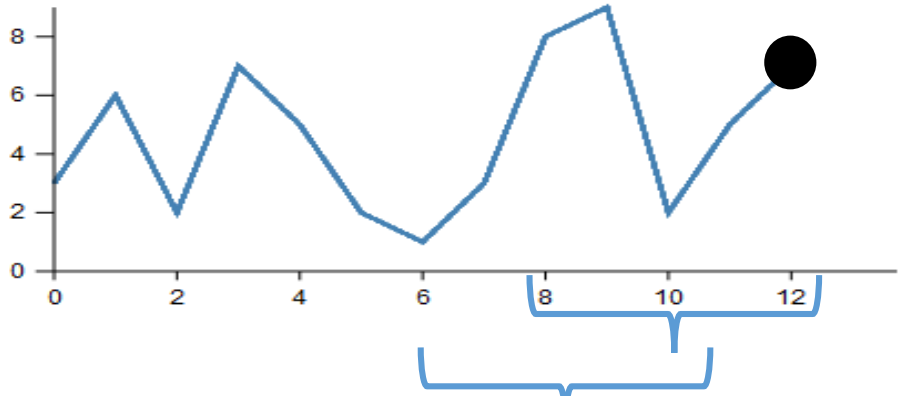


Transformations

- Standardization, normalization, feature scaling
- Categorization, binning, grouping
- Dummy variables, encoding
- Interaction effects
- Delta's, trends, windows
- Feature selection: make sure to remove "crystal balls"

Lat. 1	Long 1.	Lat. 2	Long. 2	Distance (km)	Can Walk?
48.871507	2.354350	48.872111	2.354933	2	Yes
48.872111	2.354933	44.597422	-123.248367	9059	No
48.872232	2.354211	48.872111	2.354933	5	Yes
44.597422	-123.248367	48.872232	2.354211	9056	No





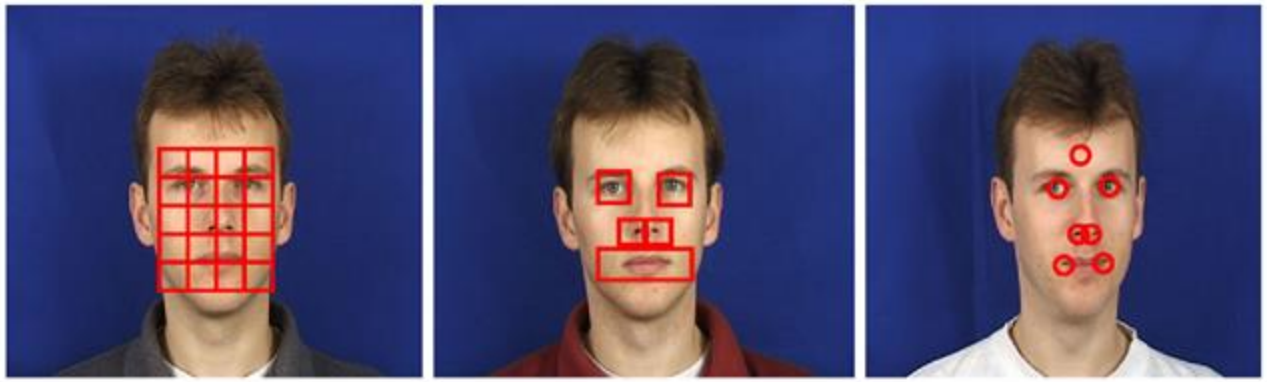
W_0 = AVERAGE(Val at 8-12)

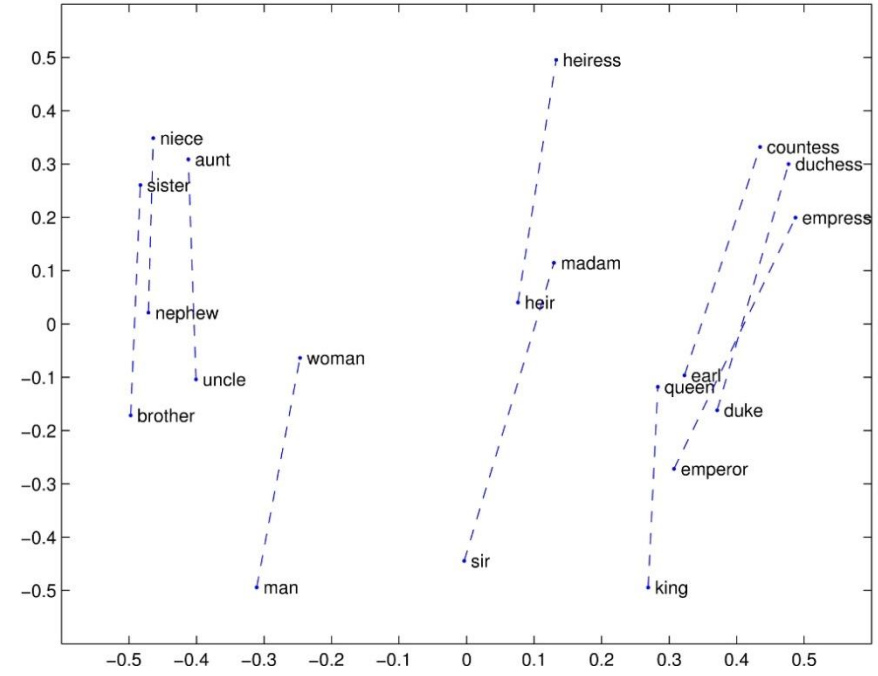
W_{-1} = AVERAGE(Val at 6-12)

Feature extraction

Squeezing features out of “difficult” (unstructured) data sources:

- Text, images, music, speech, video…
- Can be very hard and time consuming

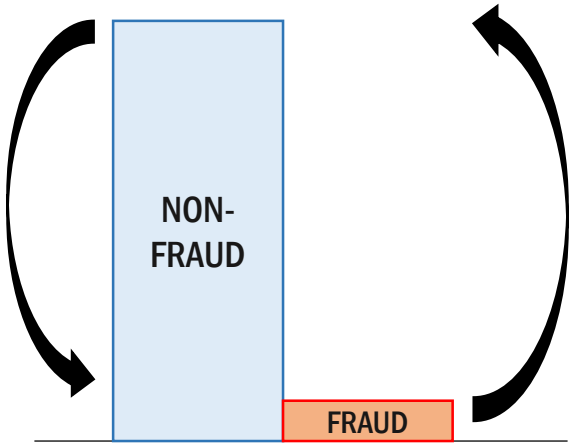




Sampling

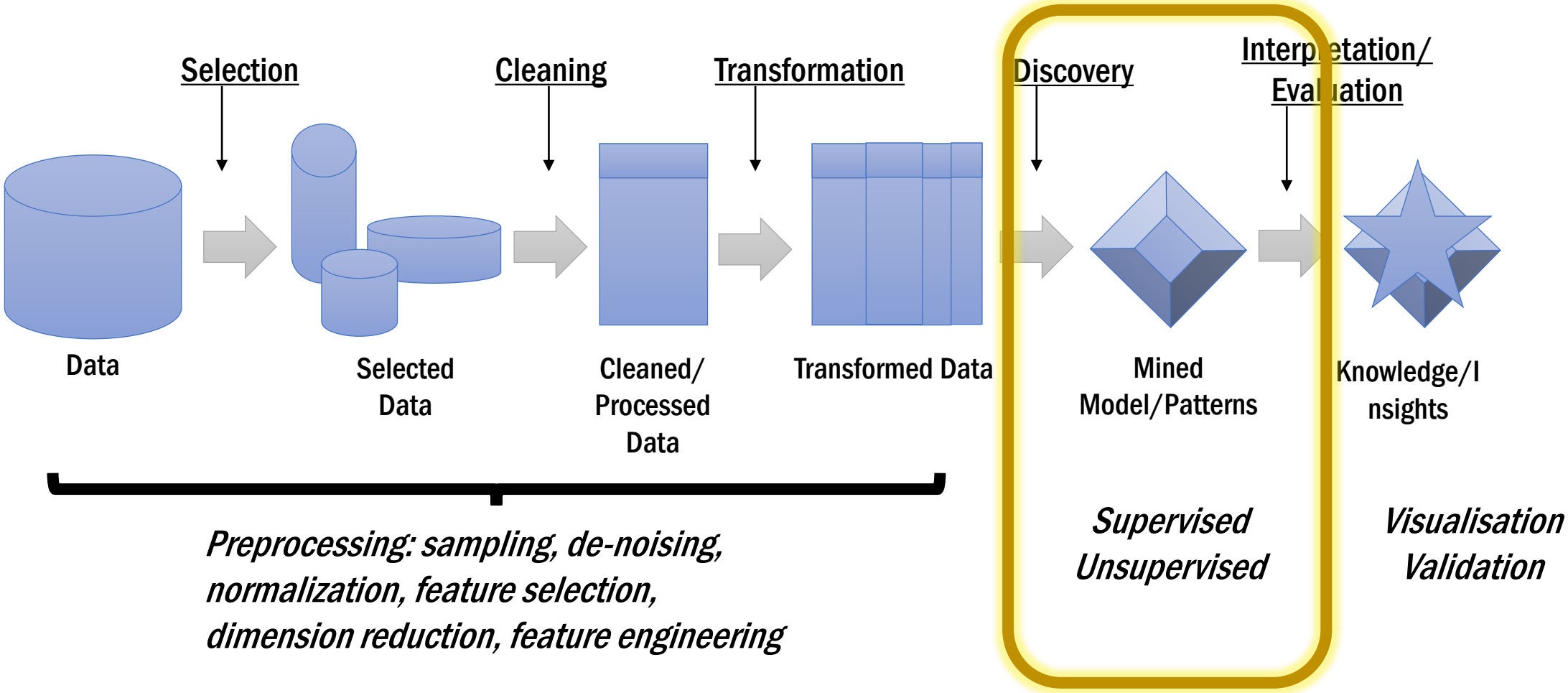
What if you have too much or highly imbalanced data (or too little)?

Down/undersampling
(faster, but we're throwing away data)



Up/oversampling
(not throwing away data, but negative effects of copying?)

The data science process



Supervised versus unsupervised

Unsupervised

- No labels/target provided, the data set “as is”
- E.g. Find groups of similar customers
  Find frequent buying patterns
- **Descriptive:** discover interesting patterns or relationships that describe the data

  Techniques: clustering, association mining

Supervised versus unsupervised

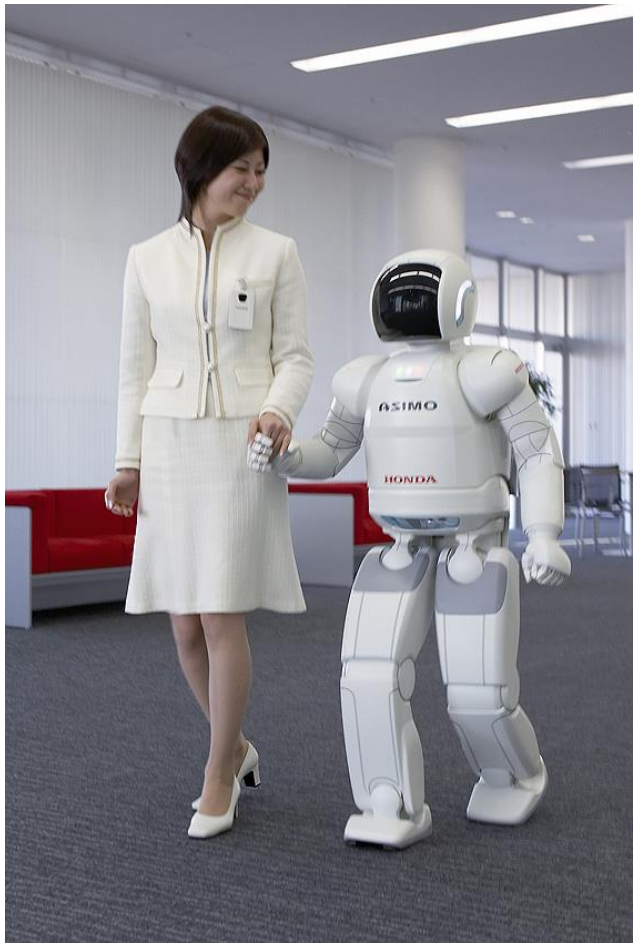
Supervised

- You have a **labelled** data set at your disposal

- E.g. List of customers with outcome *yes* if bought product and *no* if they did not
  Customers that churned versus those that did not
  Price of real estate item

- **Predictive:** predict an unknown or future value for some target variable of interest

  Techniques: classification, regression, recommender systems



Supervised versus unsupervised

- Unsupervised techniques oftentimes as a basis or starting point towards supervised ones
  - E.g. first cluster, then predict

- Other types exist as well, e.g. reinforcement learning, semi-supervised learning, etc.

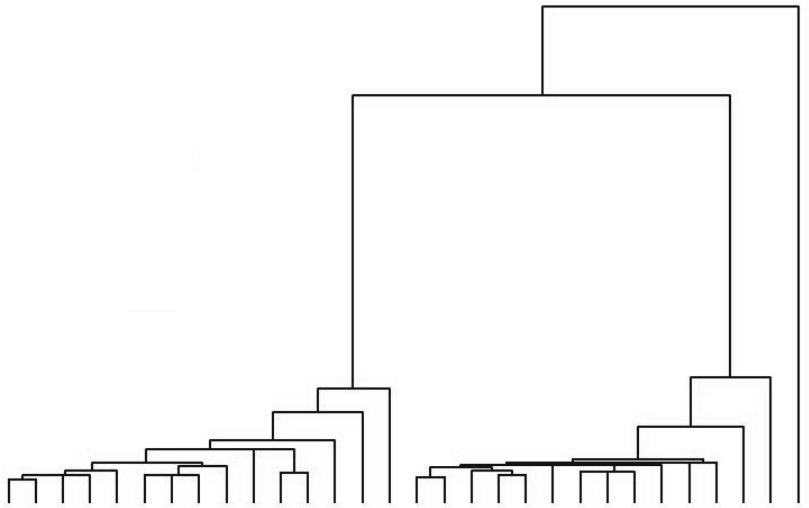
Unsupervised: clustering

- Cluster analysis or clustering is the task of **grouping a set of objects**
- In such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)
- A main task of exploratory data mining, and a common technique for statistical data analysis
- Find patterns, structure, etc.
- Sometimes the partitioning is the goal, e.g. market segmentation
- As a first step towards predictive techniques

Unsupervised: clustering

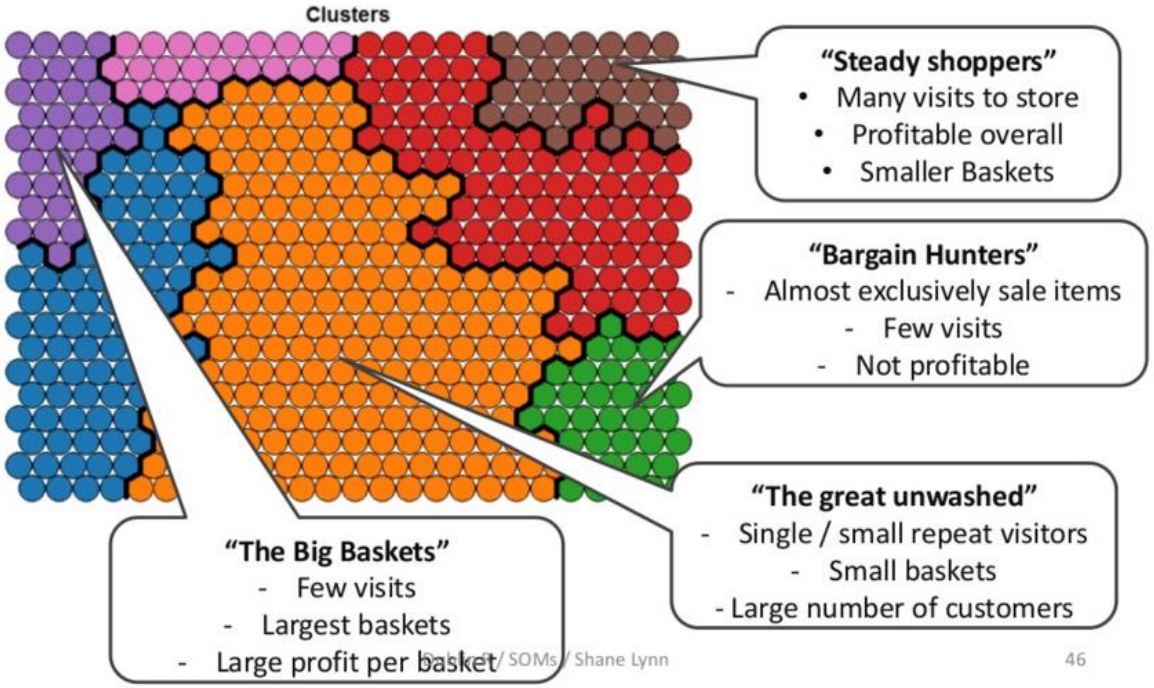
- Two types: hierarchical and partitional

- Most well-known technique: **k-means**: easy, simple, fast, possible to incorporate domain logic, but sensitive to number of clusters, initialization, non-convex patterns



Unsupervised: clustering

- Market research: customer and market segments, product positioning
- Social network analysis: recognize communities of people
- Social science: identify students, employees with similar properties
- Search: group search results
- Recommender systems: predict preferences based on user's cluster
- Crime analysis: identify areas with similar crimes
- Image segmentation and color palette detection



Unsupervised: itemset, association rule mining

- Association rule learning is a method for **discovering interesting relations between variables**

- For example, the rule *{onions, tomatoes, ketchup}* → *{burger}* found in the sales data of a supermarket can be used e.g. for promotional pricing or product placements

- “Lots of people buy lotion, but one of Pole’s colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc.”

Unsupervised: itemset, association rule mining

- Association rule learning is a method for **discovering interesting relations between variables**

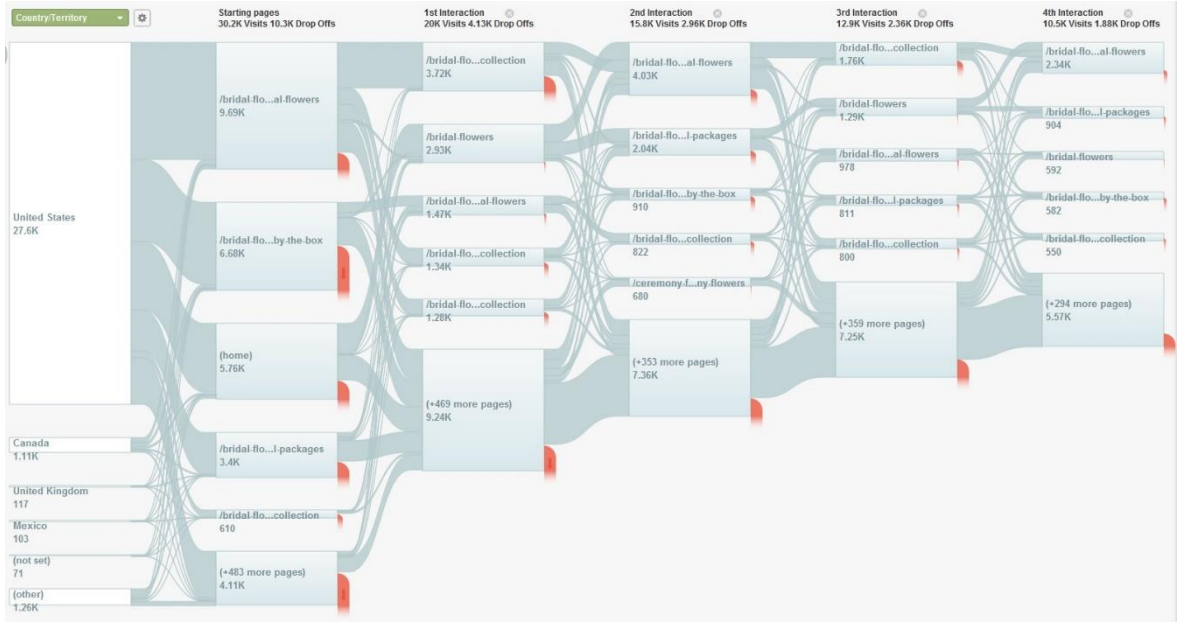
- {mortgage, checkings, fire insurance}
- {checkings, mortgage} -> {fire insurance}

- Goodness of rule assessed with metrics such as support, confidence

cus id	savings	checkings	mortgage	fire insurance	life insurance	...
101	1	1	1	0	0	0
102	0	1	1	1	1	0
103	1	1	0	1	0	1
104	0	0	0	1	1	1
105	1	1	0	1	1	1

Unsupervised: itemset, association rule mining

- Many extensions, e.g. take time into account: sequence mining

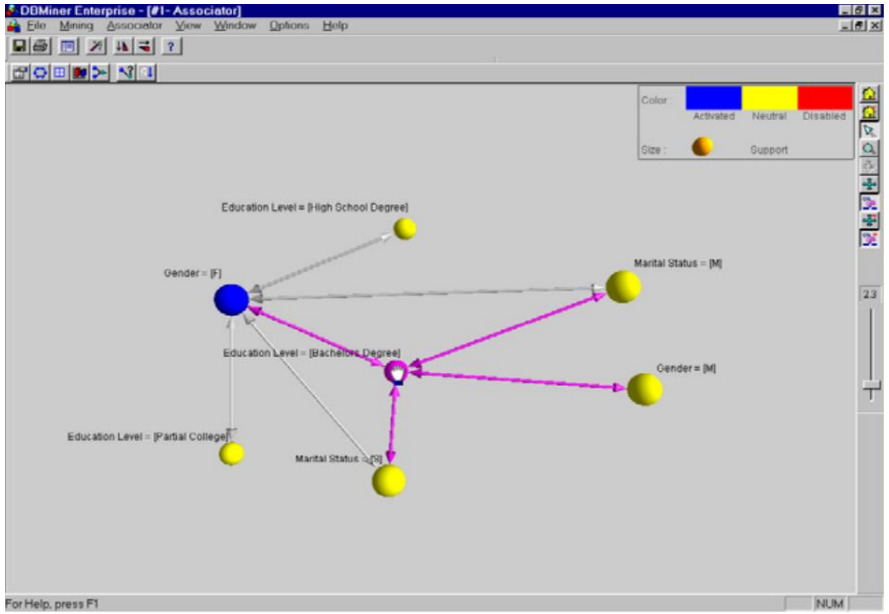


Unsupervised: itemset, association rule mining

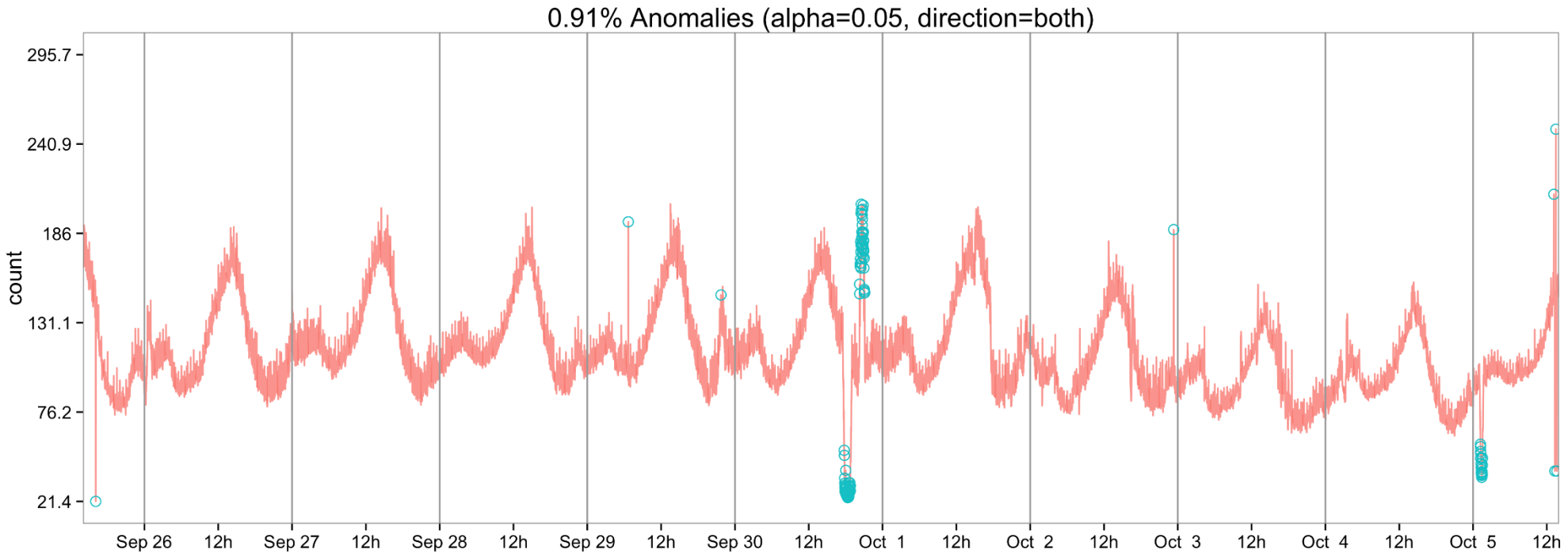
- Tune interestingness, e.g. for finding rare patterns: low support but still interesting

  E.g. people buying Rolex watches

- Context matters!



Other unsupervised techniques



Supervised

- You have a **labelled** data set at your disposal

- **Goal:** predict an unknown or future value for some target variable of interest

Algorithms that can tell you something interesting about a set of data without you having to write any custom rules, logic, code specific to the problem.

Instead of writing code, you feed data to the generic algorithm and it builds its own logic based on the data.

Supervised

As a human, your brain can approach most any situation and learn how to deal with that situation without any explicit instructions.

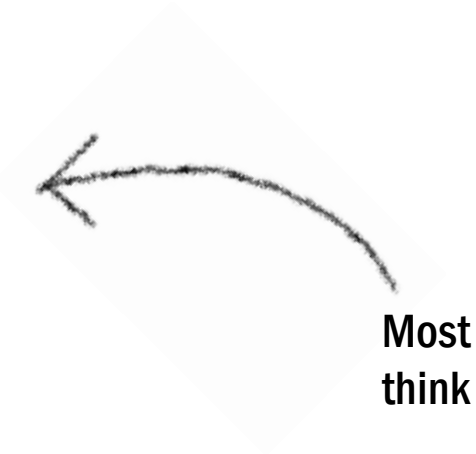
If you sell houses for a long time, you will instinctively have a "feel" for the right price for a house, the best way to market that house, etc.

Current machine learning algorithms aren't that good yet—they only work when focused a very specific, limited problem.

Maybe a better definition for "learning" in this case is "figuring out an equation or set of rules to solve a specific problem based on some example data".

Supervised

```
function estimate_house_price (num_of_bedrooms, sqm, neighborhood):

  price = 0
  price_per_sqm = 100

  if neighborhood == "woluwe":
    price_per_sqm = 400
  if neighborhood == "havenlaan":
    price_per_sqm = 20

  price = price_per_sqm * sqm

  if num_of_bedrooms == 0:
    price = price - 10000
  else:
    price = price + (num_of_bedrooms * 800)

  return price
```

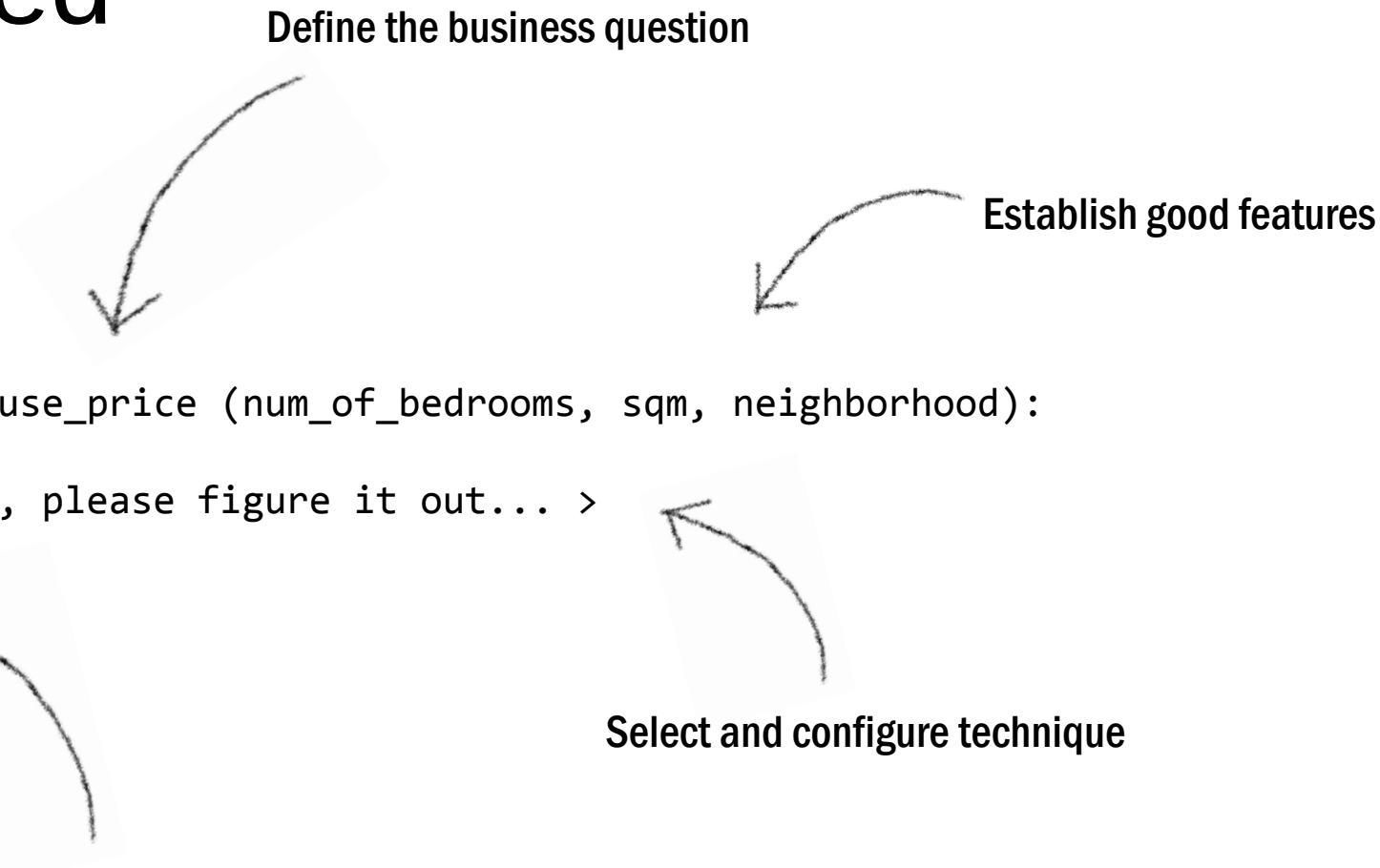


Most people focus on this when they think of “data mining”

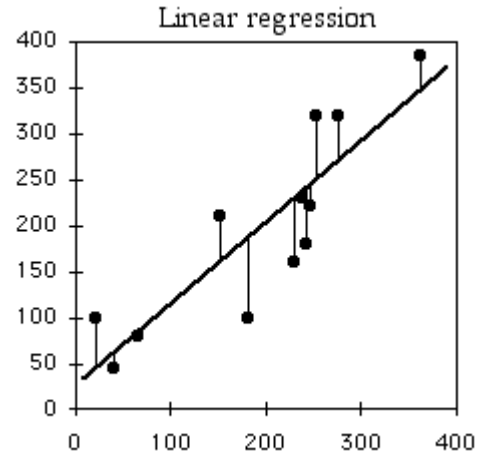
Supervised

```
function estimate_house_price (num_of_bedrooms, sqm, neighborhood):

  price = < computer, please figure it out... >

  return price
```

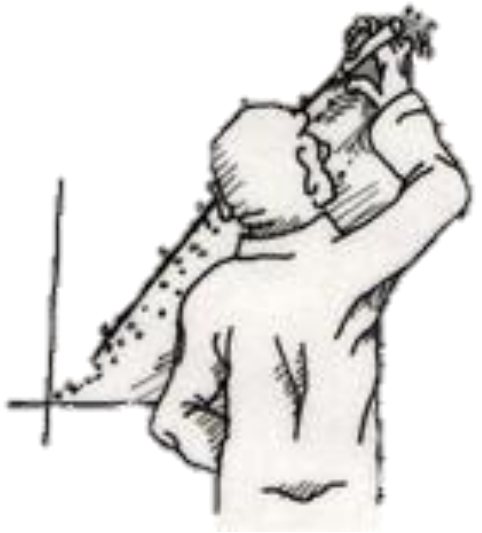
Supervised


Define the business question
Establish good features
function estimate_house_price (num_of_bedrooms, sqm, neighborhood):
price = < computer, please figure it out... >
return price
Select and configure technique
Validate and evaluate model

Supervised: regression

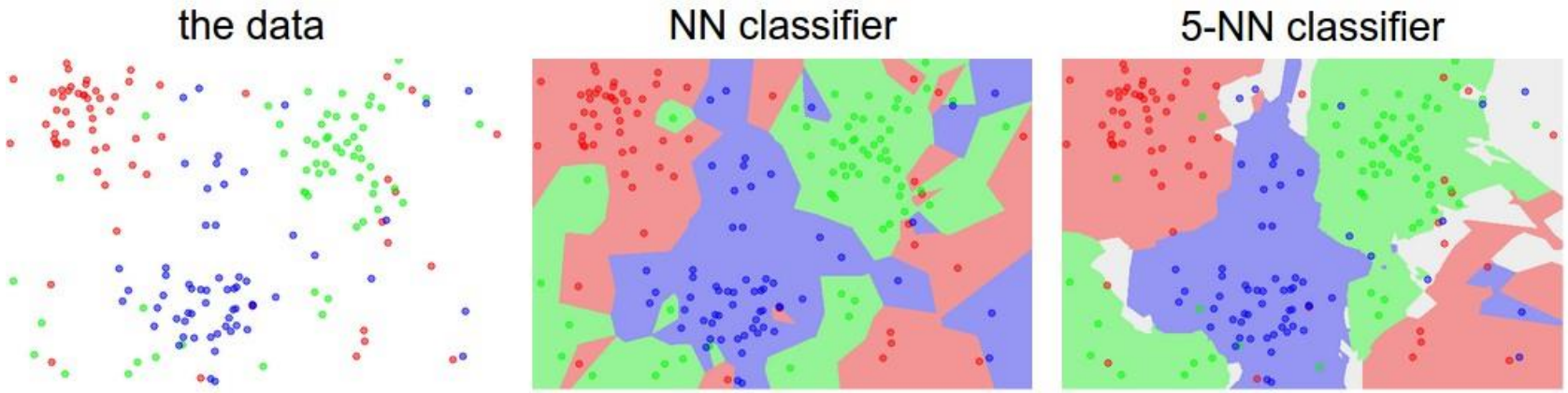


- In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables
  - E.g. university-gpa = 50 + 0.3 * highschool-gpa
- Familiar to people with statistics background
- Regression analysis is widely used for prediction and forecasting
- Strong statistical, parametric underpinnings
- Linear, logistic (for classification)



Supervised: k-NN

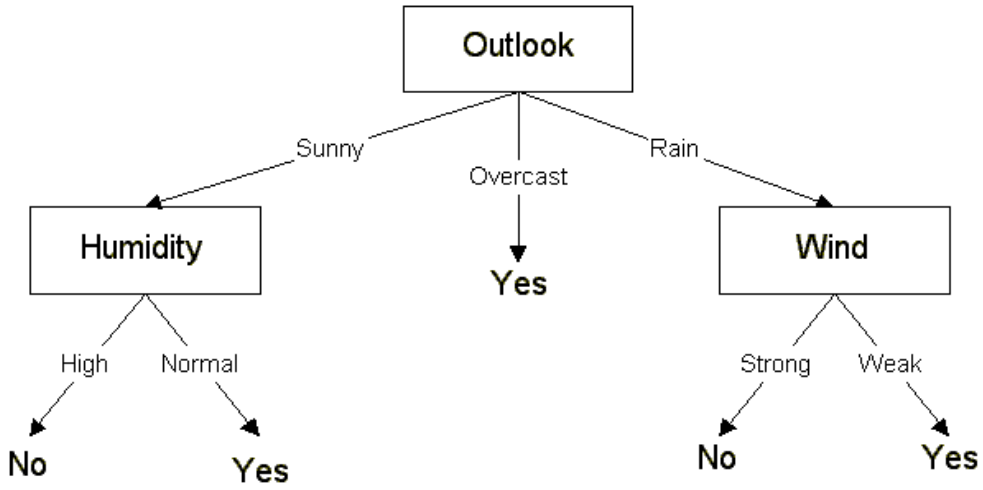
- K-nearest neighbors
- Simplest classification model imaginable
- Sensitive to amount of data



Supervised: decision trees

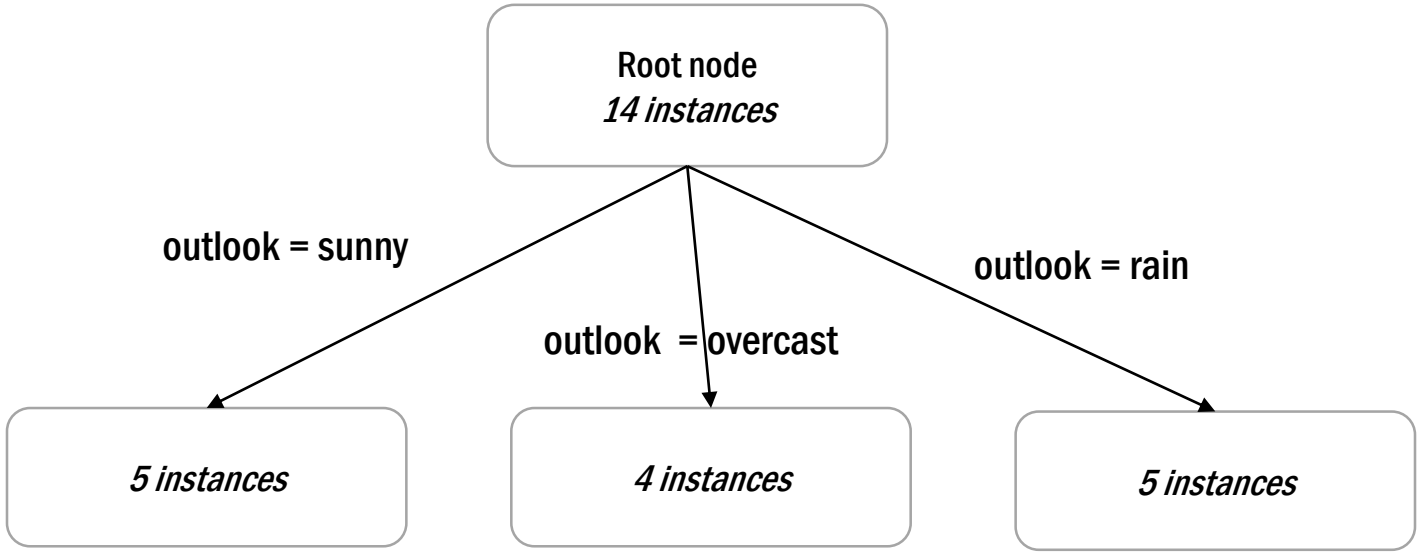
- For classification: categorical labels (classification tree)
- … but also for regression (regression tree)

- Recursively partition the data
  - How to split a node?
  - When to stop splitting?
  - How to assign a label outcome in the leaf nodes?



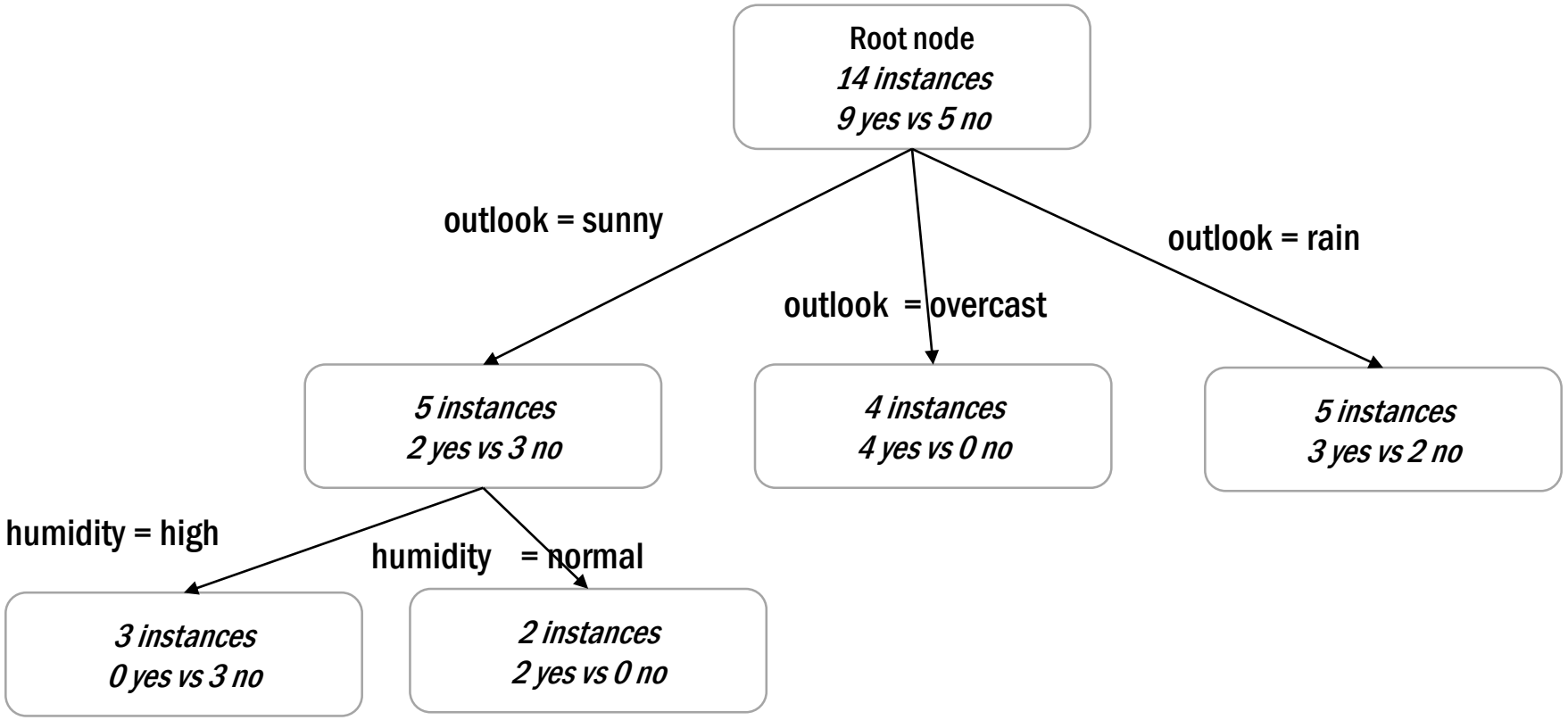
Supervised: decision trees

- Best-known technique: C4.5

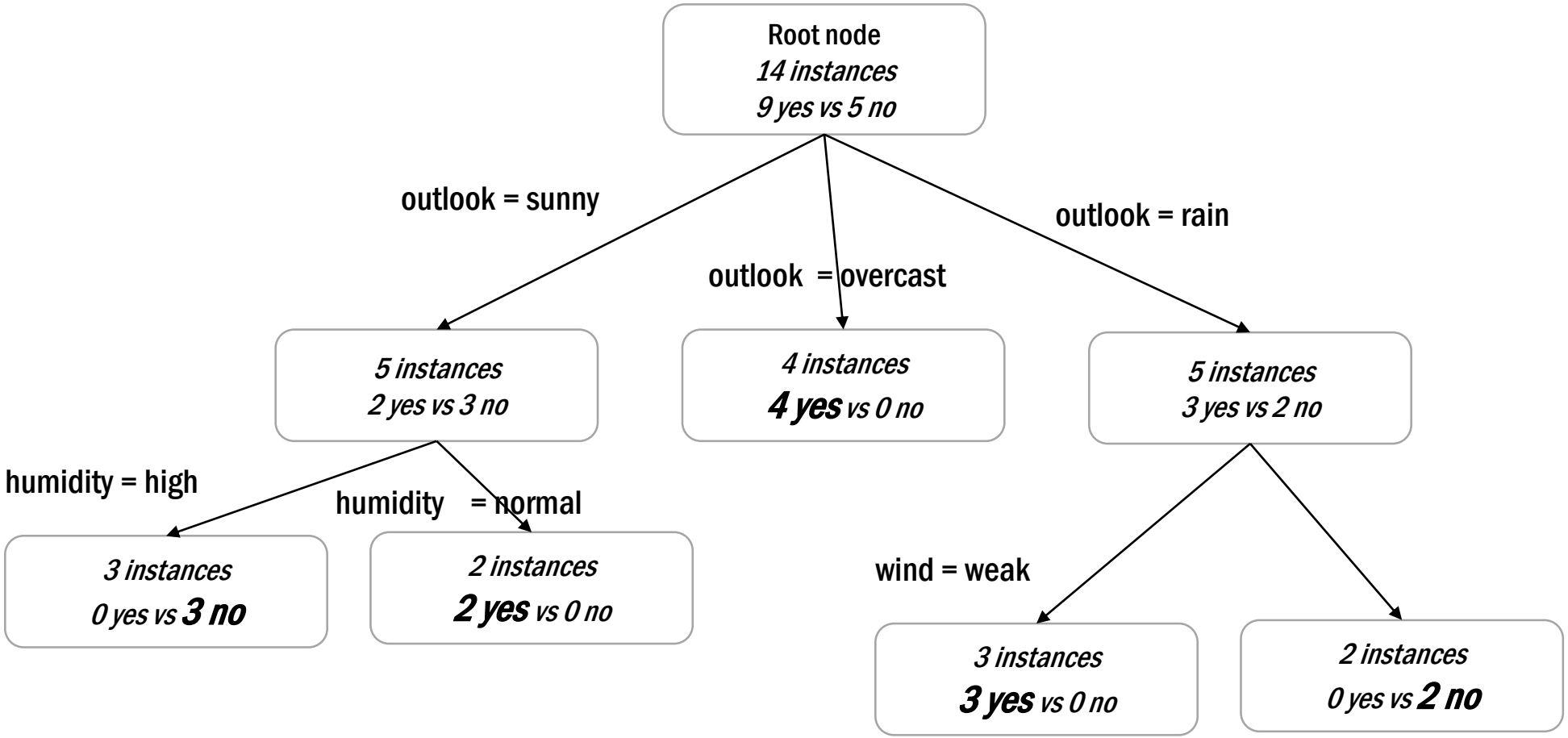


Recursion stops when every element in a subset belongs to the same class label, or there are no more attributes to be selected, or there are no instances left in the subset

Supervised: decision trees



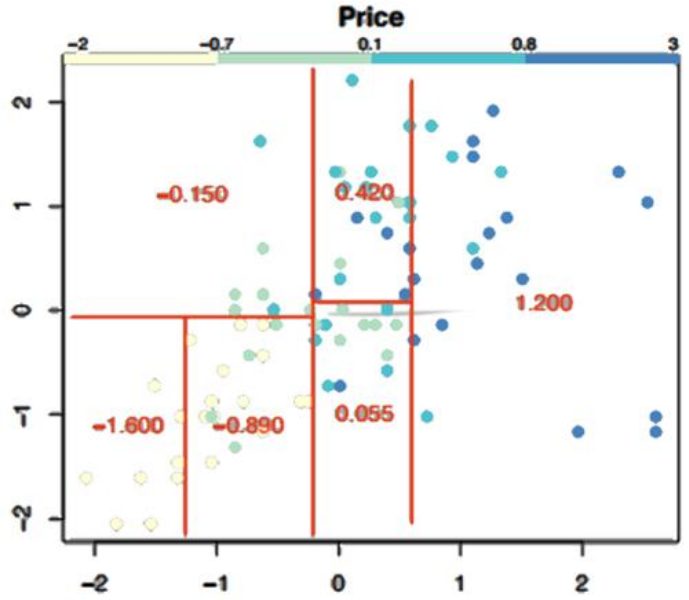
Supervised: decision trees

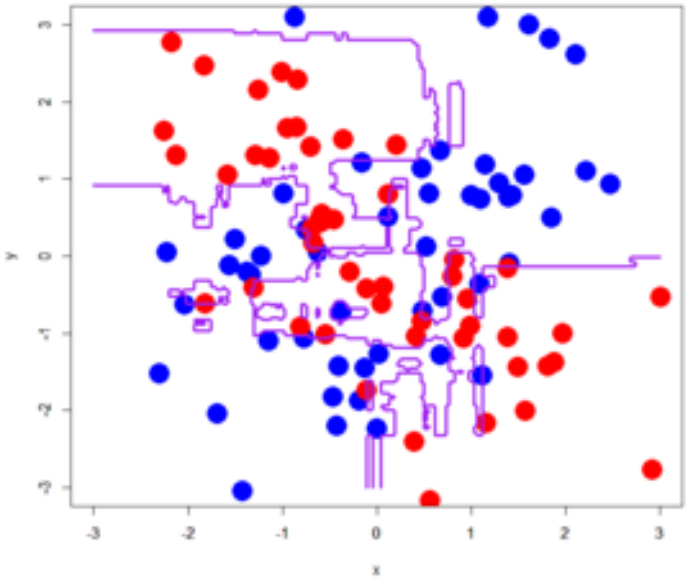


Supervised: decision trees

- Very robust technique
- Simple to understand, requires very little data preparation
- Can handle both numerical and categorical variables
- White-box model
- Can also be used for regression
- Performs well on large datasets

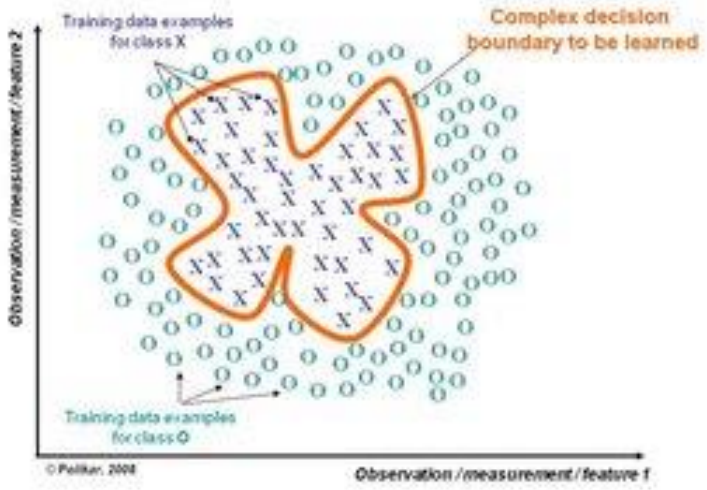
- But: prone to overfitting, hence common usage of pruning or early stopping

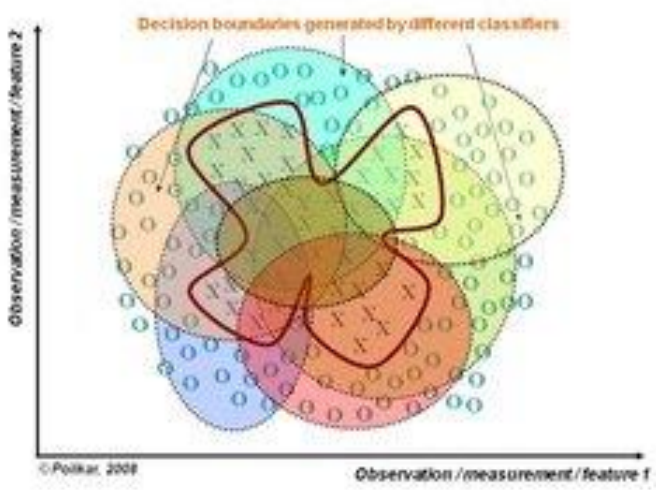




Supervised: random forests

- Best known example of an “ensemble method”
- Operate by constructing a multiple decision trees at training time and outputting the class that is the majority vote of the classes (classification) or mean prediction (regression) of the individual trees at prediction-time





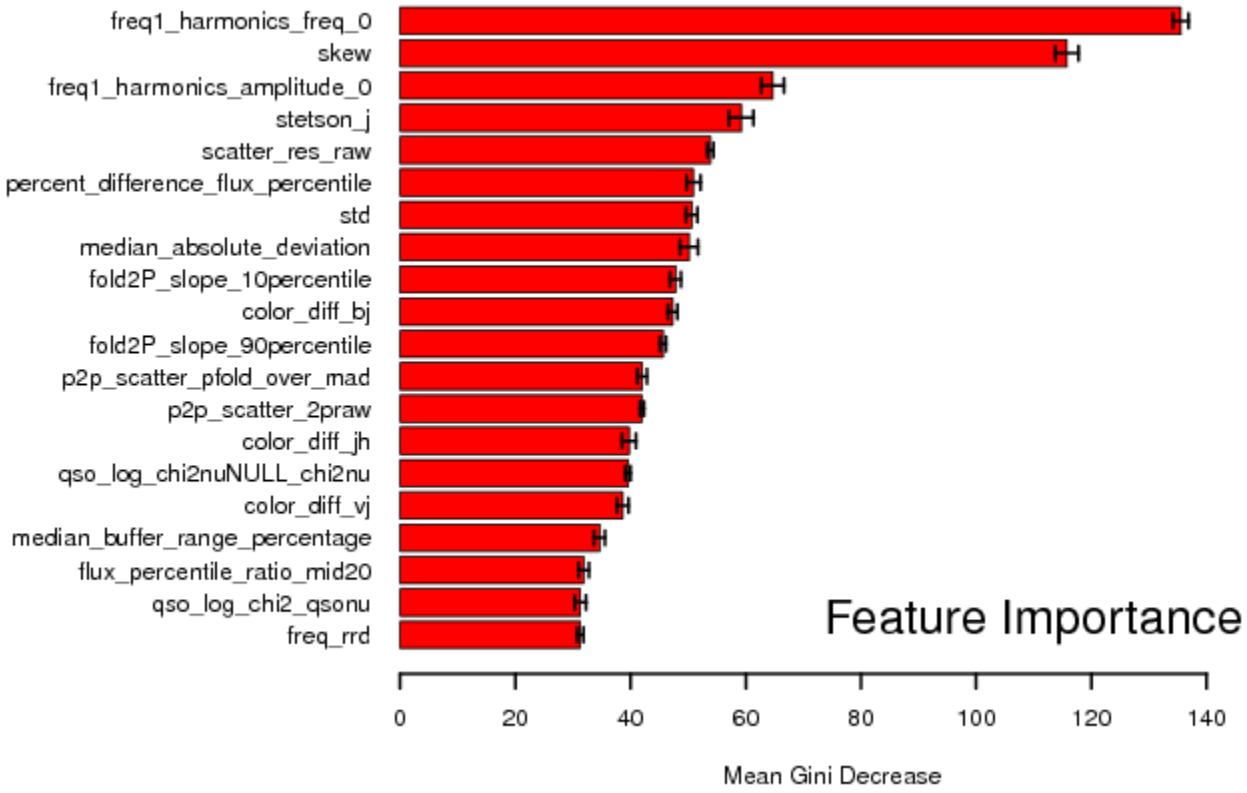
Supervised: random forests

- One part of randomness comes from bootstrapping each decision tree, i.e. each decision tree sees a random sample of the training data
- However, random forests use an additional piece of randomness, i.e. to select the allowed attributes to split on
- Random forests are extremely easy to use, don’t require (that) much configuration or preprocessing, good at avoiding overfitting (by design)
- However… how to explain 100 trees vs. 1? Random forests are a black box…

Supervised: random forests

How to explain 100 trees vs. 1?

- **Feature importance**: you can examine which variables are working best/worst in each tree, i.e. when a certain tree uses one variable and another doesn't, you can compare the value lost or gained from the inclusion/exclusion of that variable

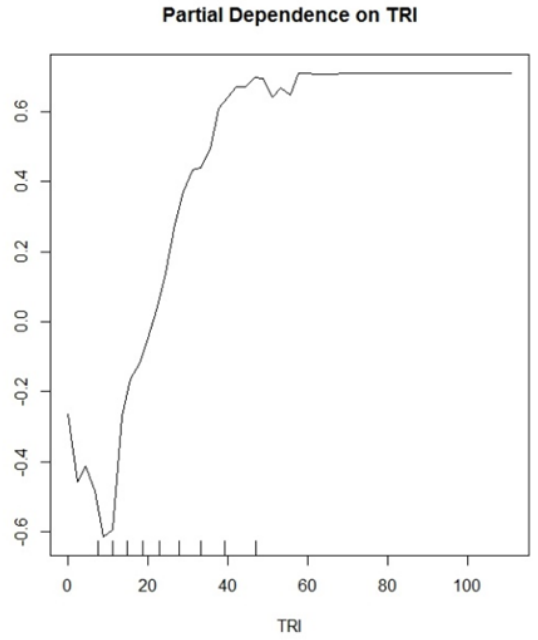


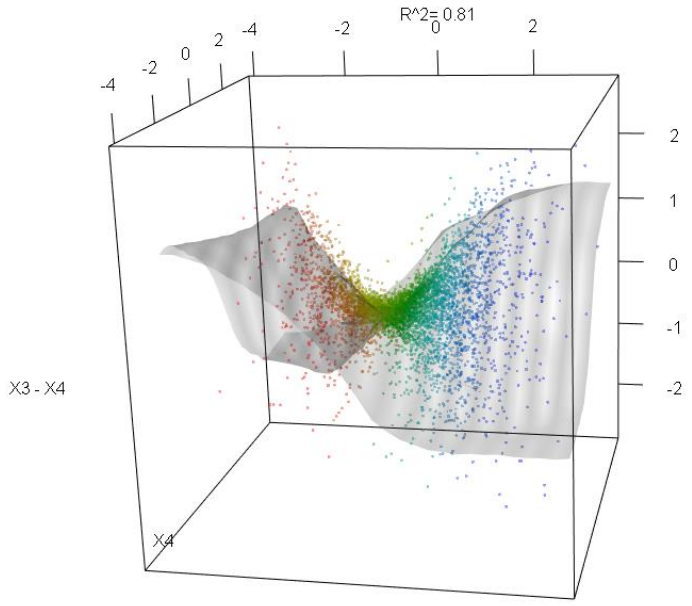
Supervised: random forests

How to explain 100 trees vs. 1?

- **Partial dependencies**: each point on the partial dependence plot is the average vote percentage in favor of the “yes“-class across all observations, given a fixed level of the feature under observation

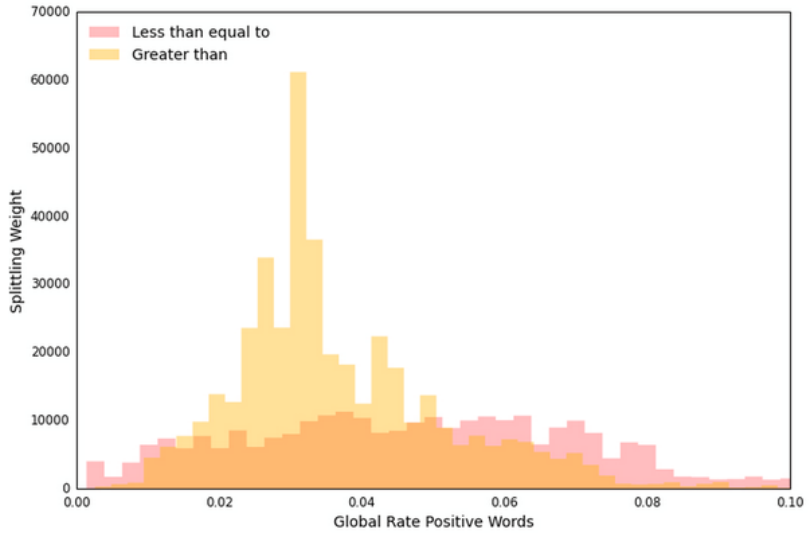
  Good for inspecting main effects, but not interactions





Supervised: random forests

How to explain 100 trees vs. 1?



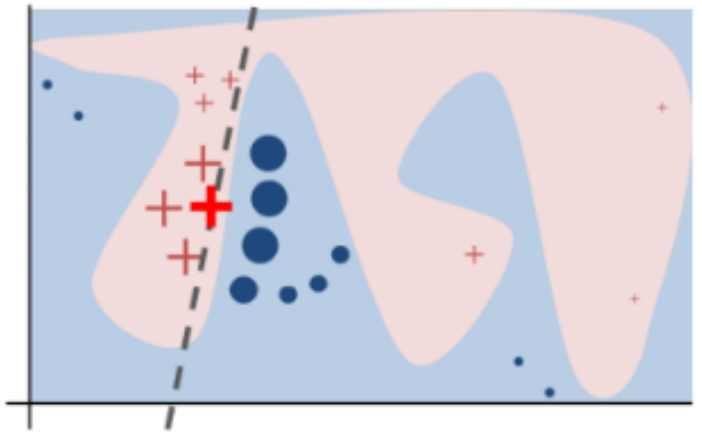
- **Most-used-variable**: “feed” an instance through every tree in the forest and tally which variable was used more often

- **Decision paths**: for a forest, the prediction is simply the average of the bias terms plus the average contribution of each feature

- **Split analysis**: “feed” an instance through every tree in the forest, but now also keep track of the splitting points per variable

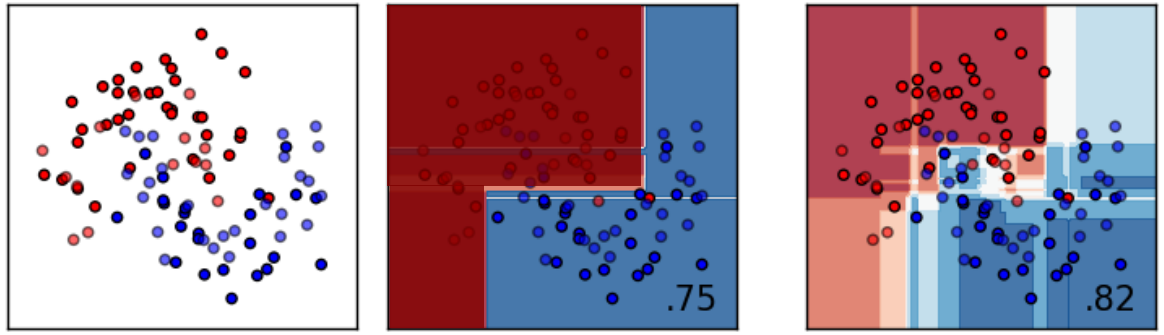
Supervised: random forests

How to explain 100 trees vs. 1?



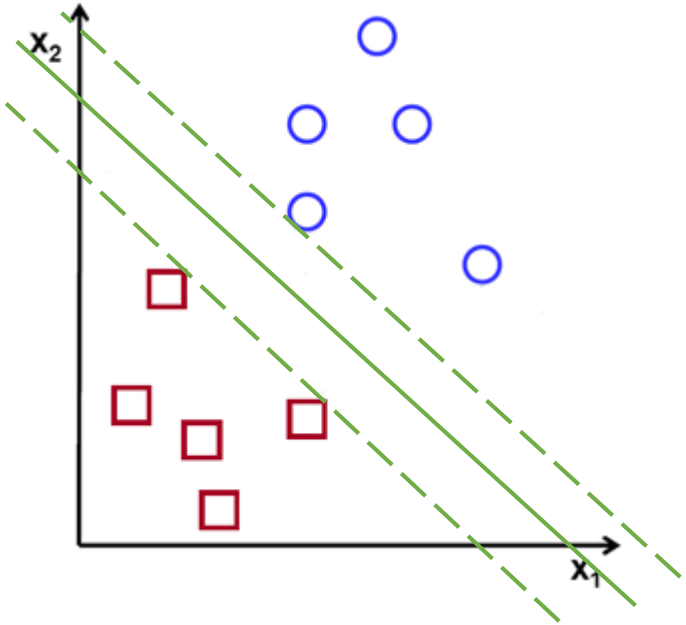
- **Construct simple local model**: to explain the classification of a particular instance

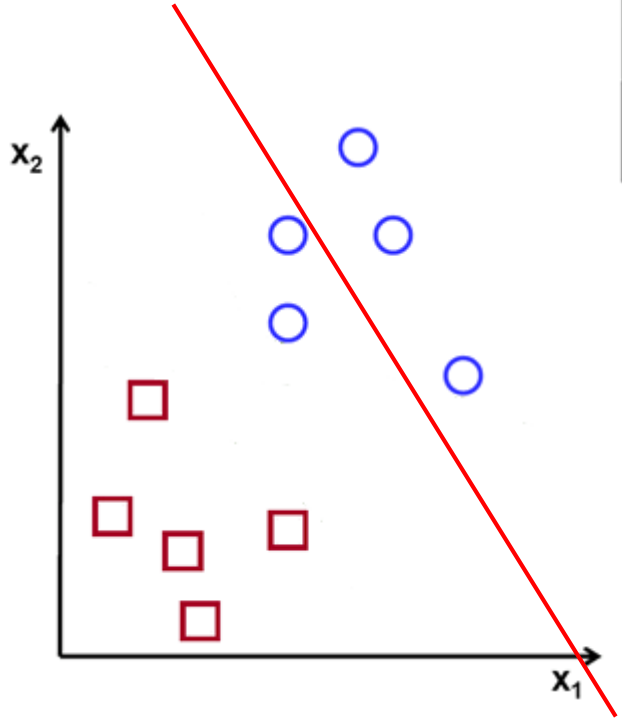
- **Construct simple global model**: i.e. use the black-box model, but explain it using a simpler, white box model (a single tree)



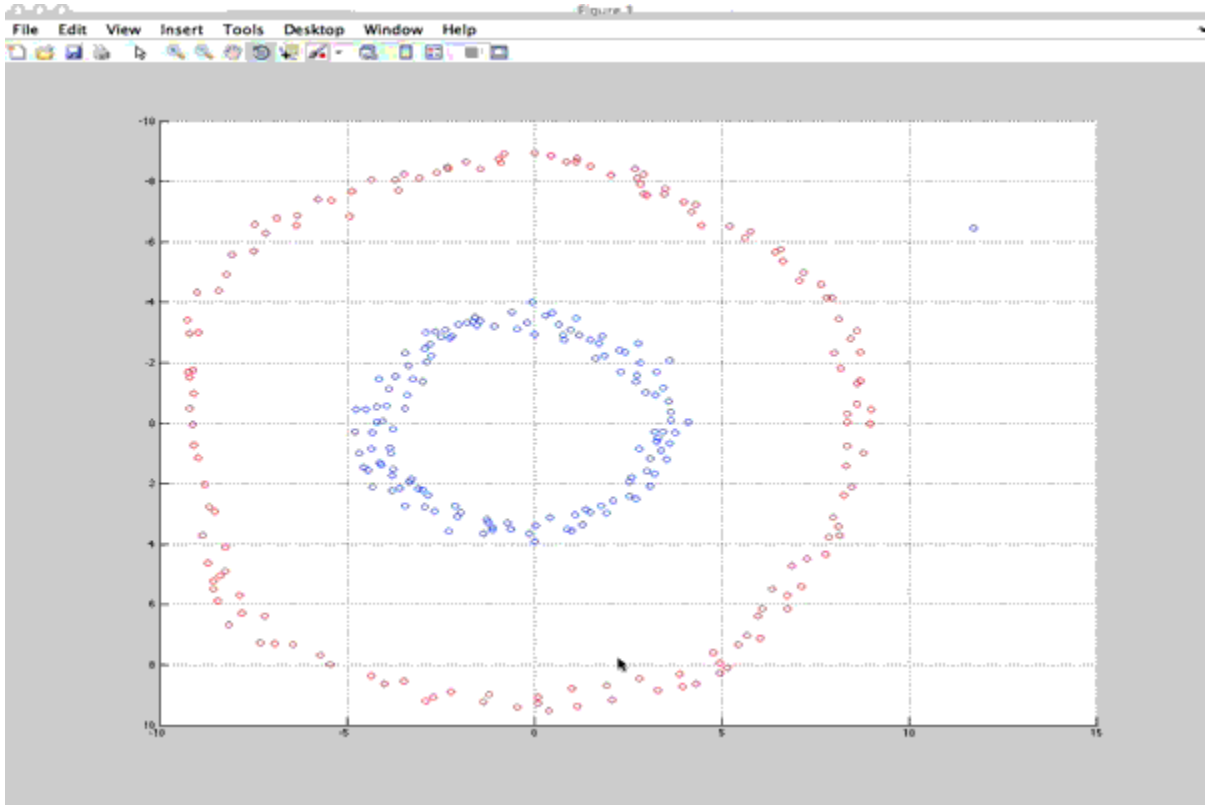
Other supervised techniques: support vector machines

- Trying to find a good hyperplane
- What if there is no good enough line possible?
- … our optimization breaks down
- … unless: we try a new perspective!



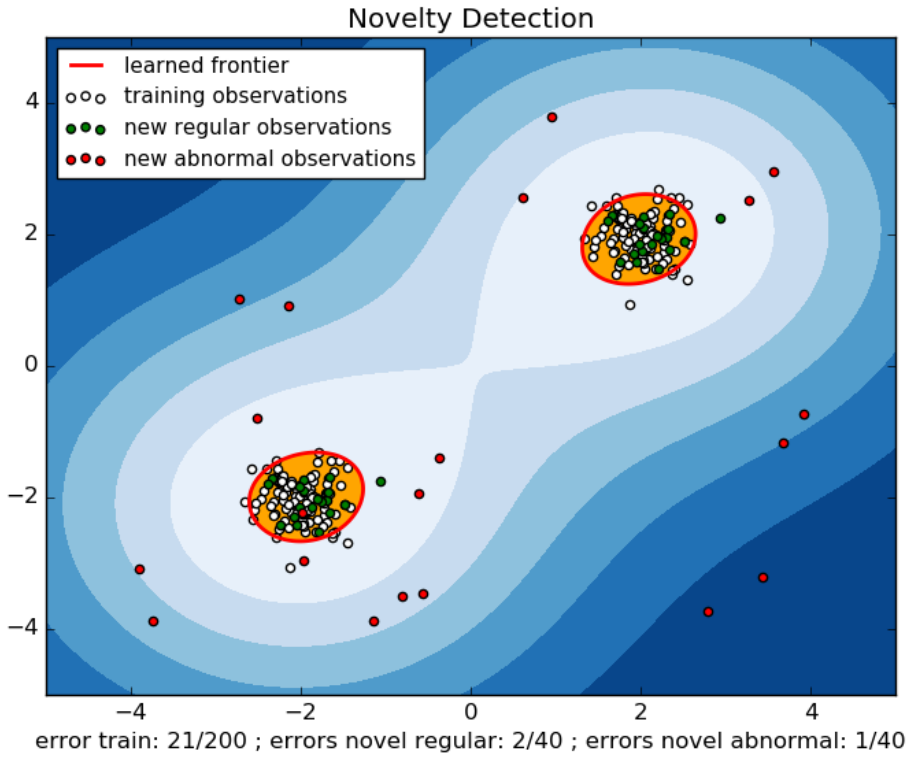


Other supervised techniques: support vector machines

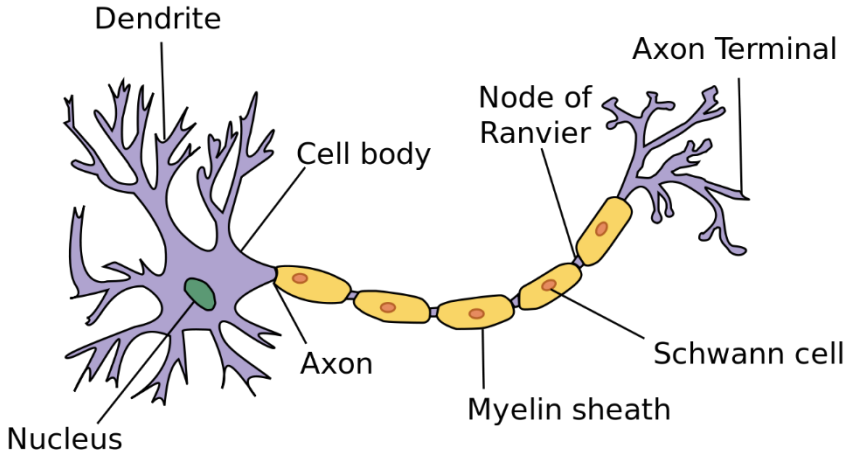


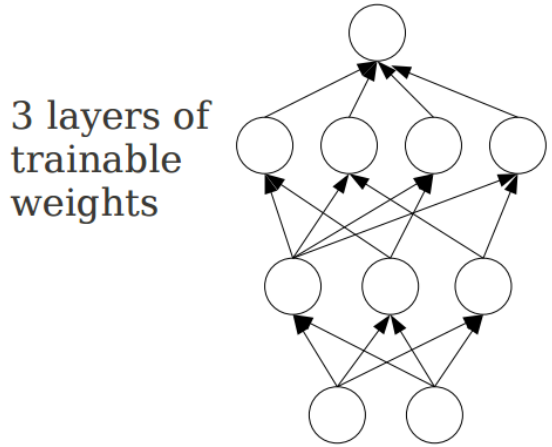
Other supervised techniques: support vector machines

- For classification and regression too
- For anomaly detection (unsupervised)
- For datasets with many features
- Beautiful convex, global optimization problem
- … but their compute and storage requirements increase rapidly with the number of training vectors
- … requires extensive configuration



Other supervised techniques: artificial neural networks





Very loosely based on how (we think) the human brain works

- A collection of software “neurons” are created and connected together, allowing them to send messages to each other
- Next, the network is asked to solve a problem, which it attempts to do over and over, each time strengthening the connections that lead to success and diminishing those that lead to failure

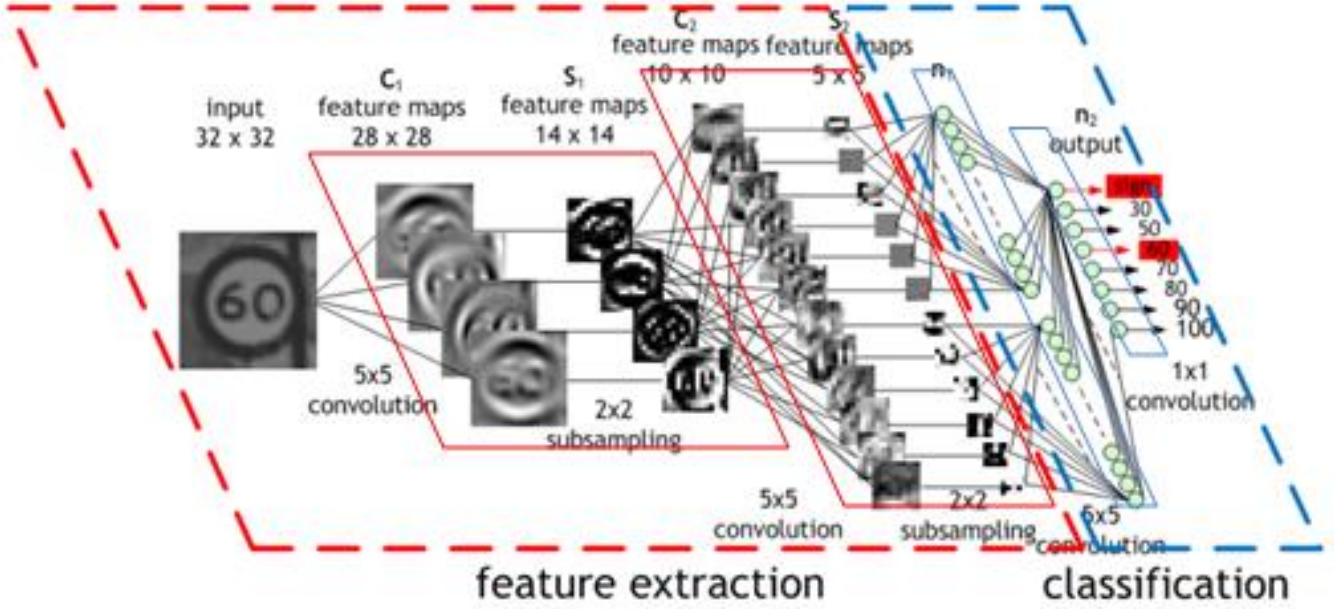
Other supervised techniques: artificial neural networks

Powerful, but…

- How many hidden layers?
- How many nodes in each layer?
- Which activation function(s)?
- Which layout?
- Which learning rate? (Too small goes too slowly, too high might cause overshooting.) Adaptive or not?
- Which feature transformers?

Risk of overfitting!

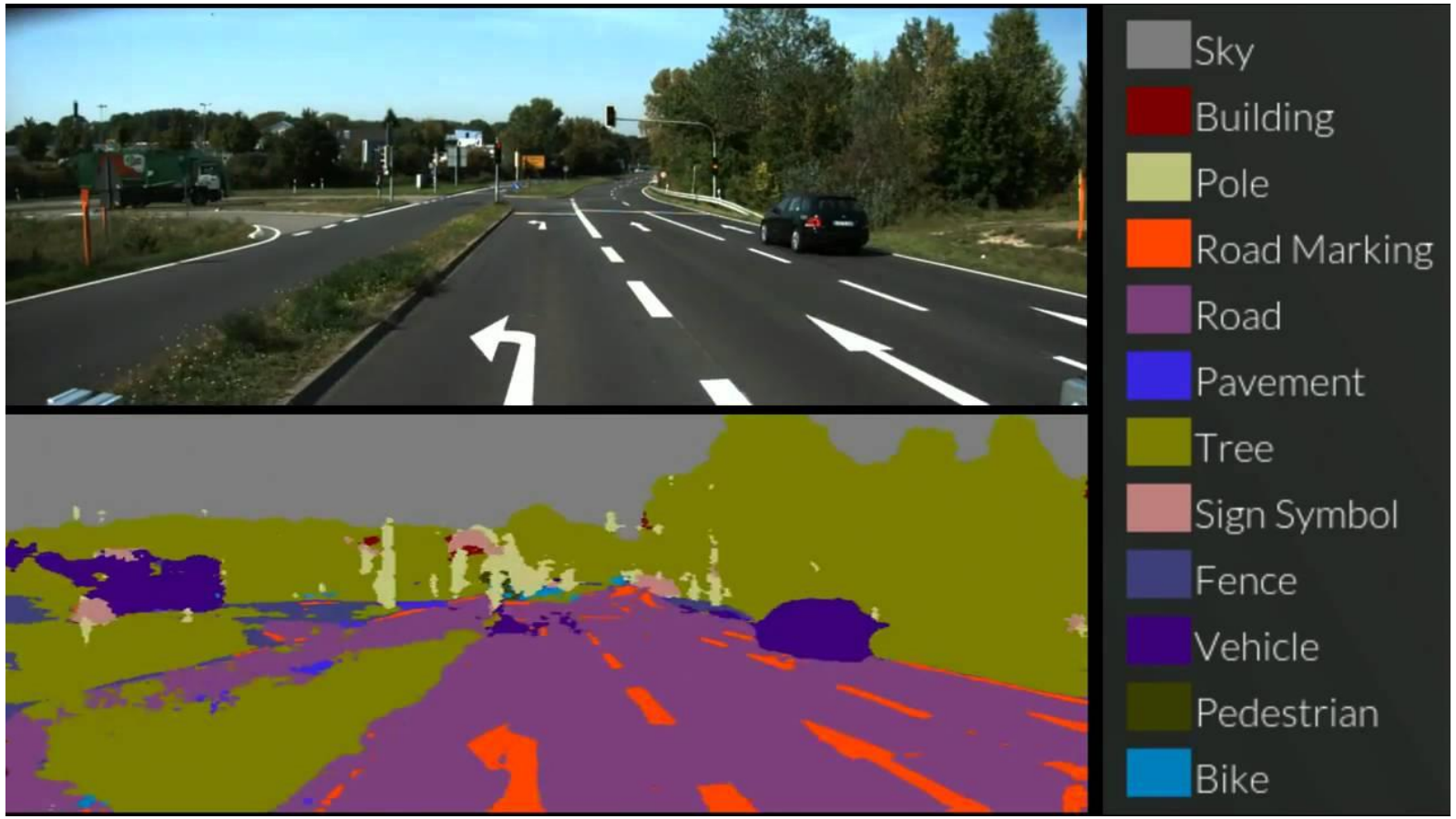
Other supervised techniques: deep learning



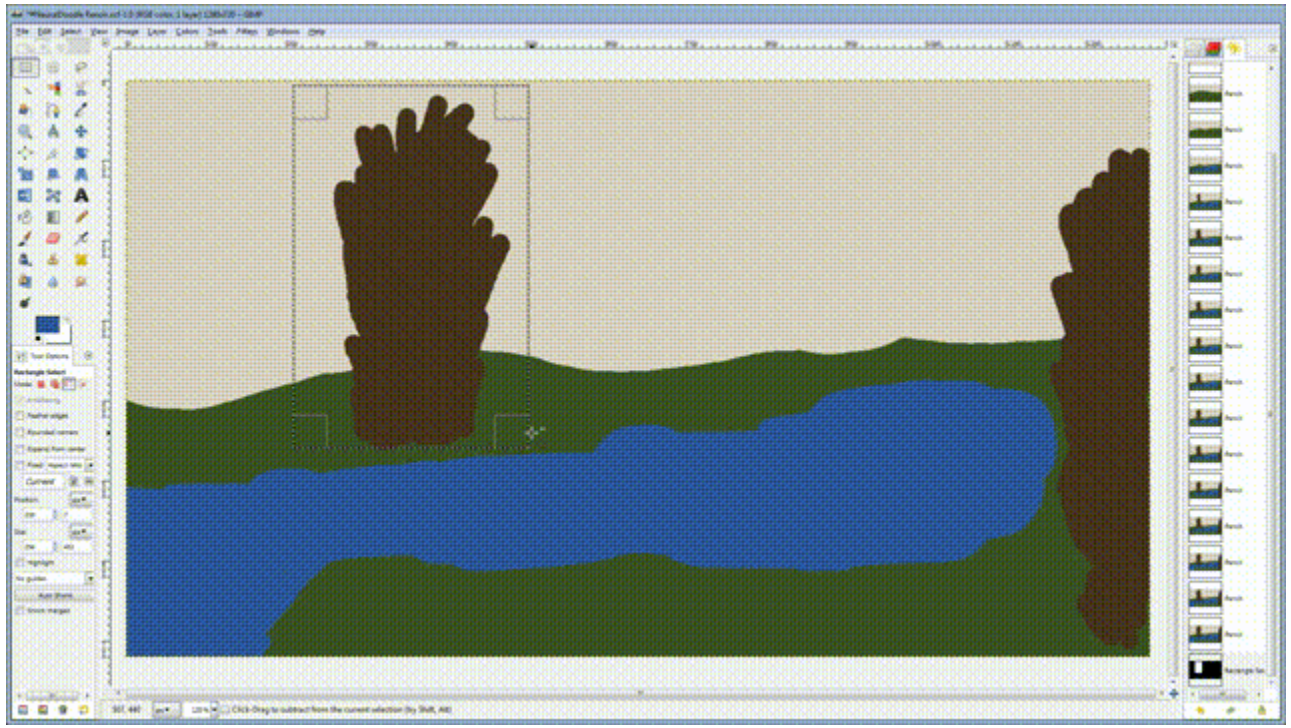
Other supervised techniques: deep learning

Other supervised techniques: deep learning

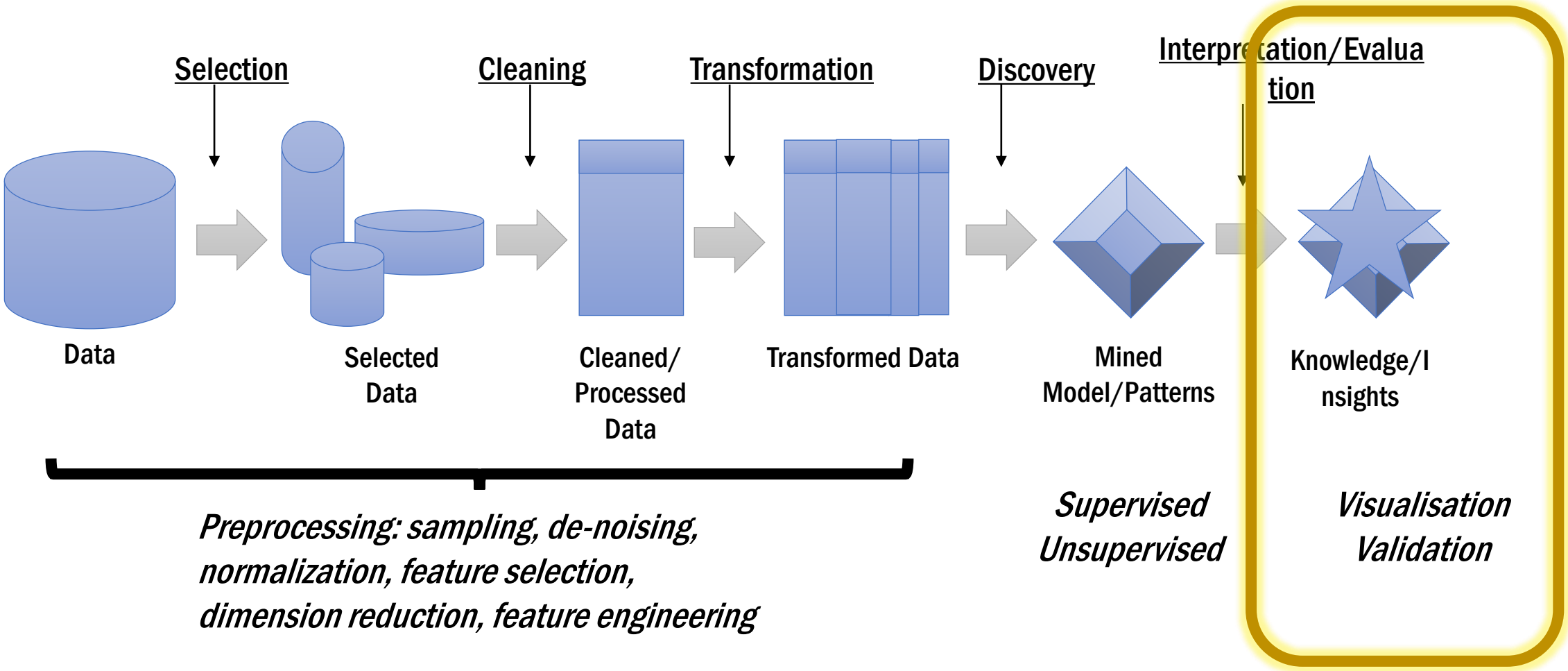
Other supervised techniques: deep learning



Other supervised techniques: deep learning



The data science process



Model validation and evaluation

What to validate?

- Model specification (e.g. selection of variables, definition of business question)
- Model quantification (e.g. estimation of coefficients, lay-out of decision tree)
- Model performance: the predictive ability of the model

Types of validity

- Apparent (own sample)
- Internal (own population)
- External (other population)
- Model performance on train, test, validation set

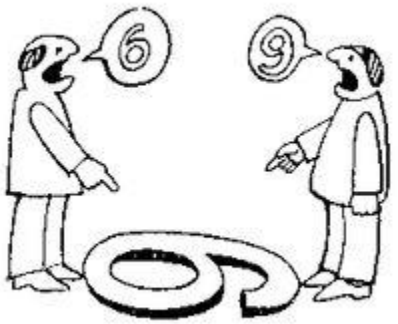
Solve the business problem

This requires a thorough business knowledge and understanding of the problem to be addressed before any analysis can start

Some example kick-off questions are: how do we define (what is it?), measure (how to see it?) and manage (what to do with it?) fraud, churn, a sale…?

Often many ways to define a problem:

- E.g. predict which customers will reach gold status next year
- Classification on the current group at time t?
- Classification on the difference t -> t+1?
- Regression on the value determining the gold status threshold?
- Time series forecasting on the value determining the gold status threshold?
- How much is enough? How much do you want to know?



Model validation and evaluation

True Label	Prediction	Predicted Label	Correct?
no	0.11	no	Correct
no	0.20	no	Correct
yes	0.85	yes	Correct
yes	0.84	yes	Correct
yes	0.80	yes	Correct
no	0.65	yes	Incorrect
yes	0.44	no	Incorrect
no	0.10	no	Correct
yes	0.32	no	Incorrect
yes	0.87	yes	Correct
yes	0.61	yes	Correct
yes	0.60	yes	Correct
yes	0.78	yes	Correct
no	0.61	yes	Incorrect

Threshold 0.50

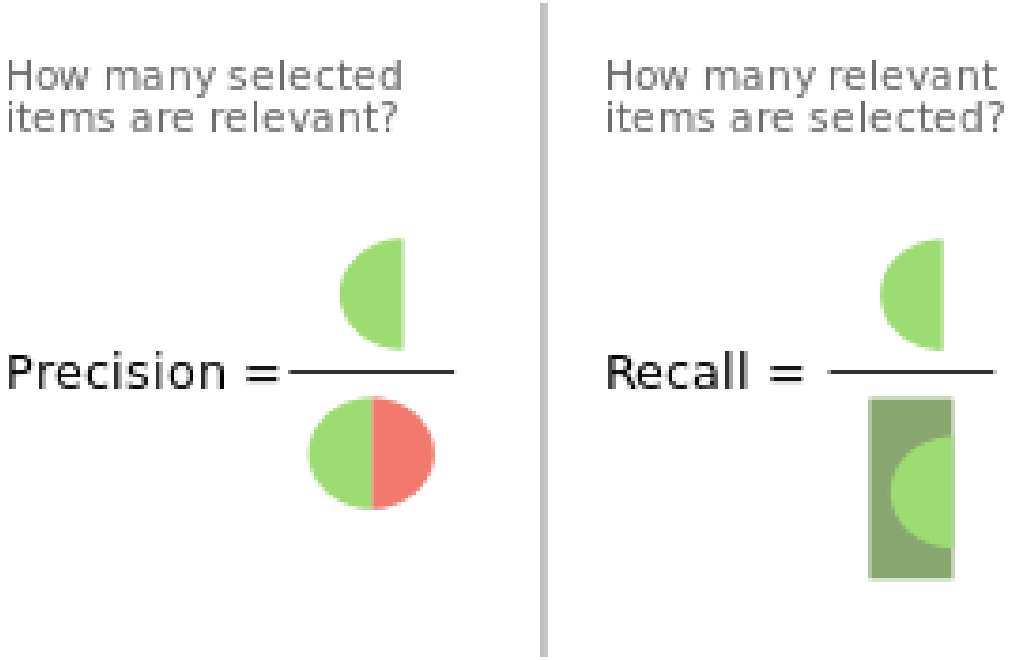
Confusion matrix:

		Prediction	
Reference		*no*	*yes*
	no	3 tn	2 fp
	yes	2 fn	7 tp

Model validation and evaluation

Accuracy: how many predictions did we get right?

Recall vs. **precision**:

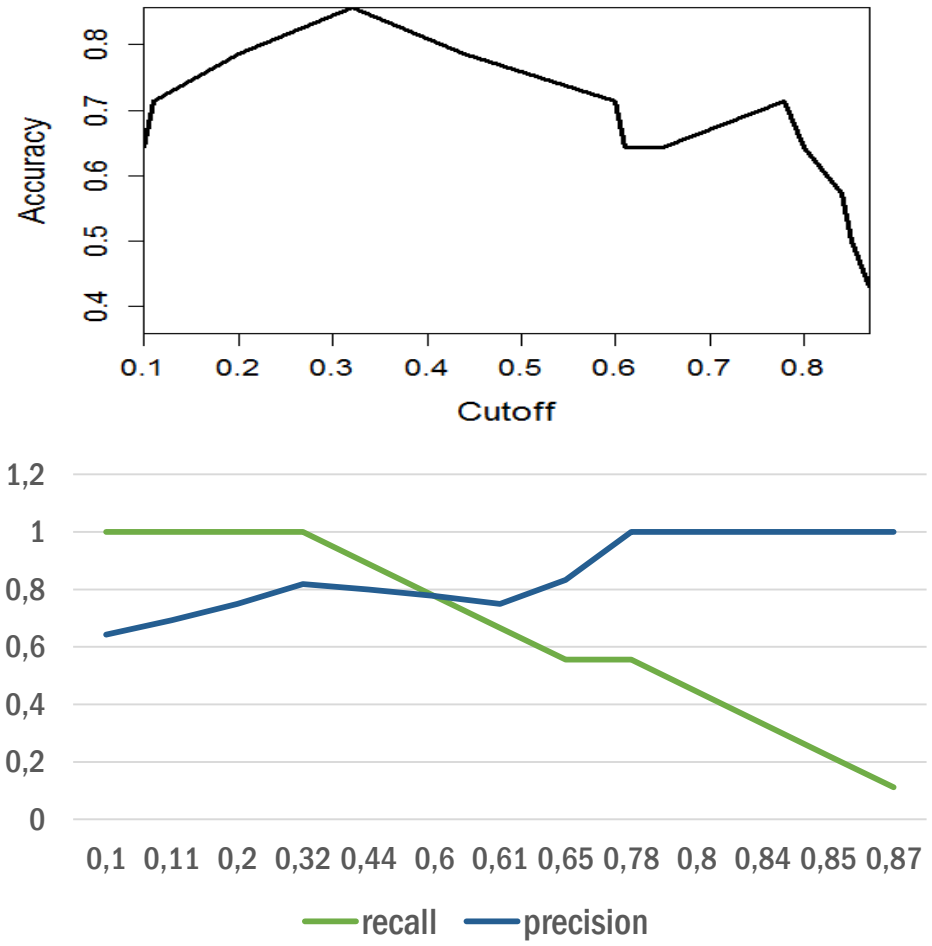


Reference	Prediction: no	Prediction: yes
no	3 tn	2 fp
yes	2 fn	7 tp

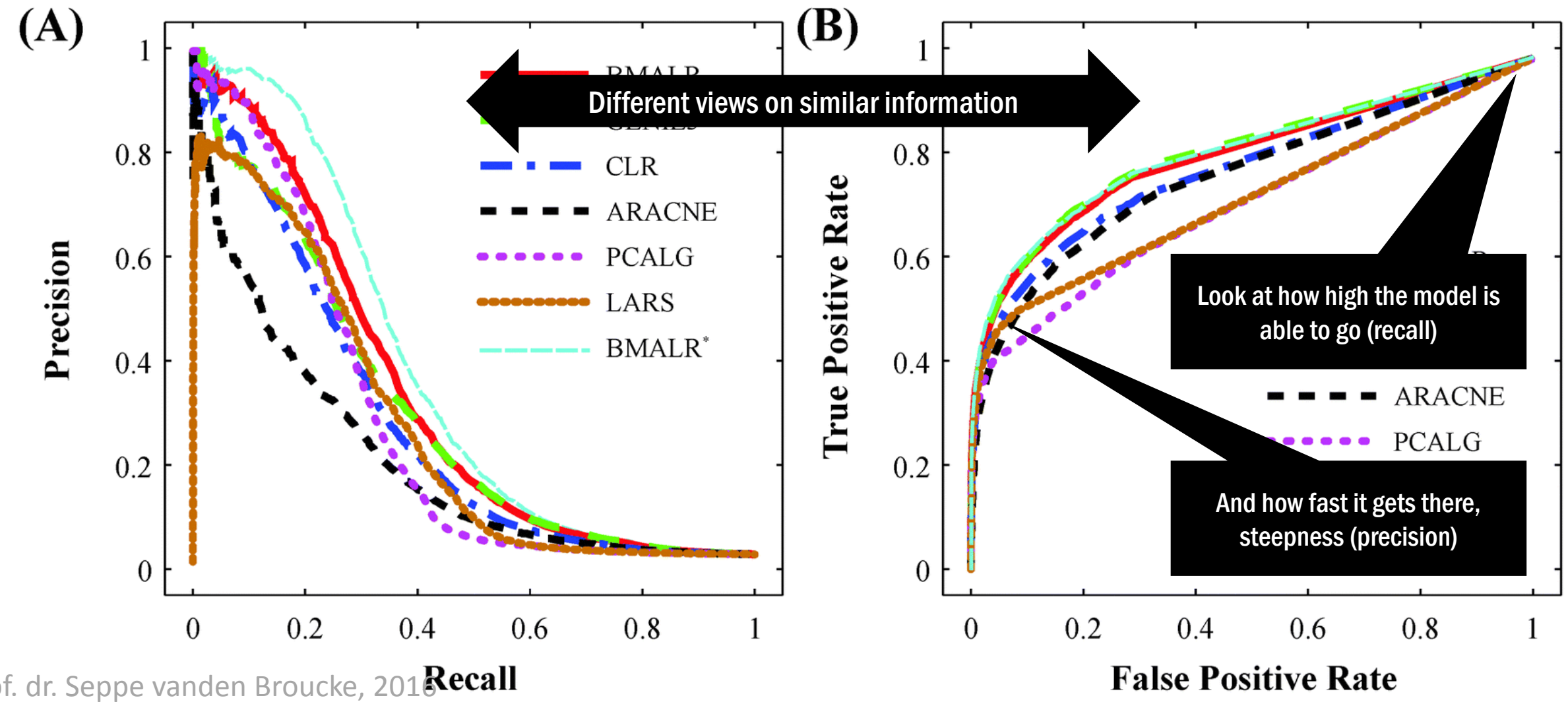
Model validation and evaluation

True Label	Prediction	Predicted Label	Correct?
no	0.11	no	Correct
no	0.20	no	Correct
yes	0.85	yes	Correct
yes	0.84	yes	Correct
yes	0.80	yes	Correct
no	0.65	yes	Incorrect
yes	0.44	no	Incorrect
no	0.10	no	Correct
yes	0.32	no	Incorrect
yes	0.87	yes	Correct
yes	0.61	yes	Correct
yes	0.60	yes	Correct
yes	0.78	yes	Correct
no	0.61	yes	Incorrect

Best threshold



Model validation and evaluation



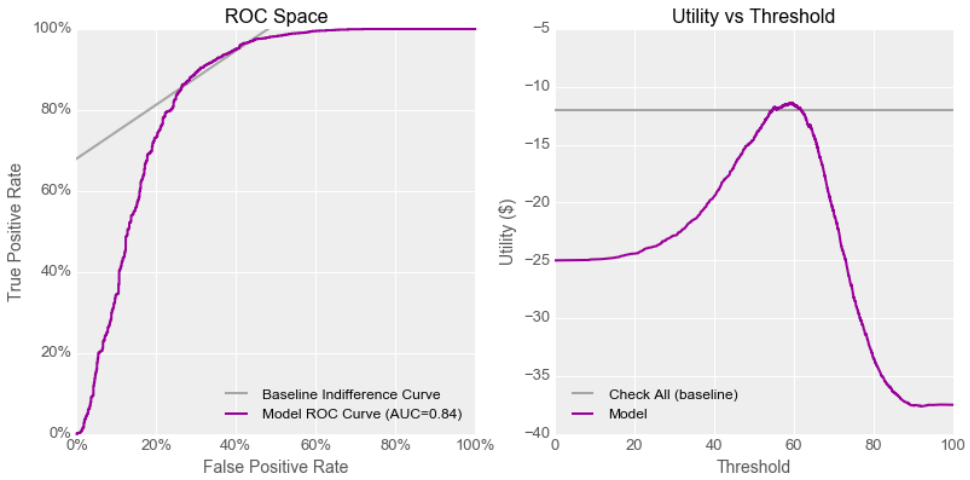
Model validation and evaluation

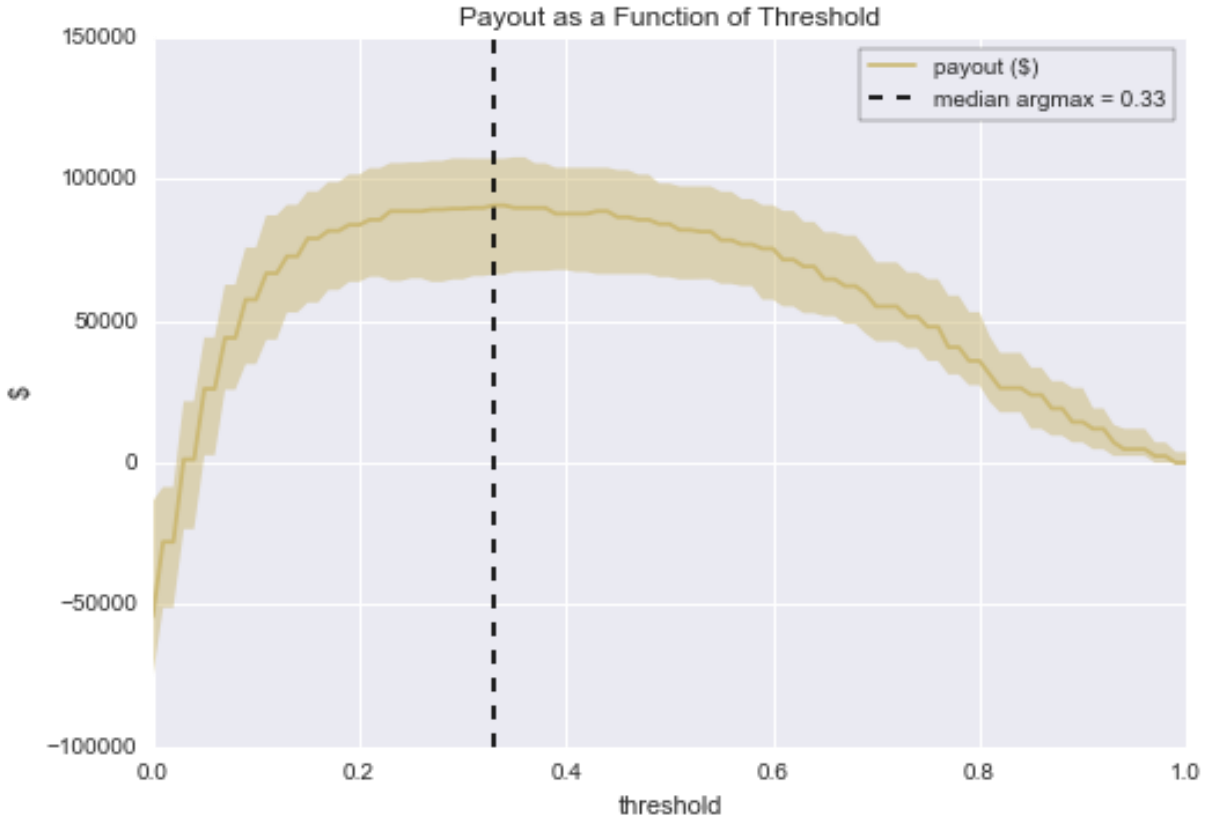
Lift (ratio of a model to a random guess)

- E.g. assume a random model handing out random probabilities
- Take the top n (top 100)
- See how many of them were indeed “yes”, e.g. 10 / 100
- Now do the same for your model, gives e.g. 80 / 100
- Lift of your model over random is 80 / 10 = 8
- Depends on how many n (in general, getting more hits is more difficult in a shorter list), and apriori class distribution between “no” and “yes” instances

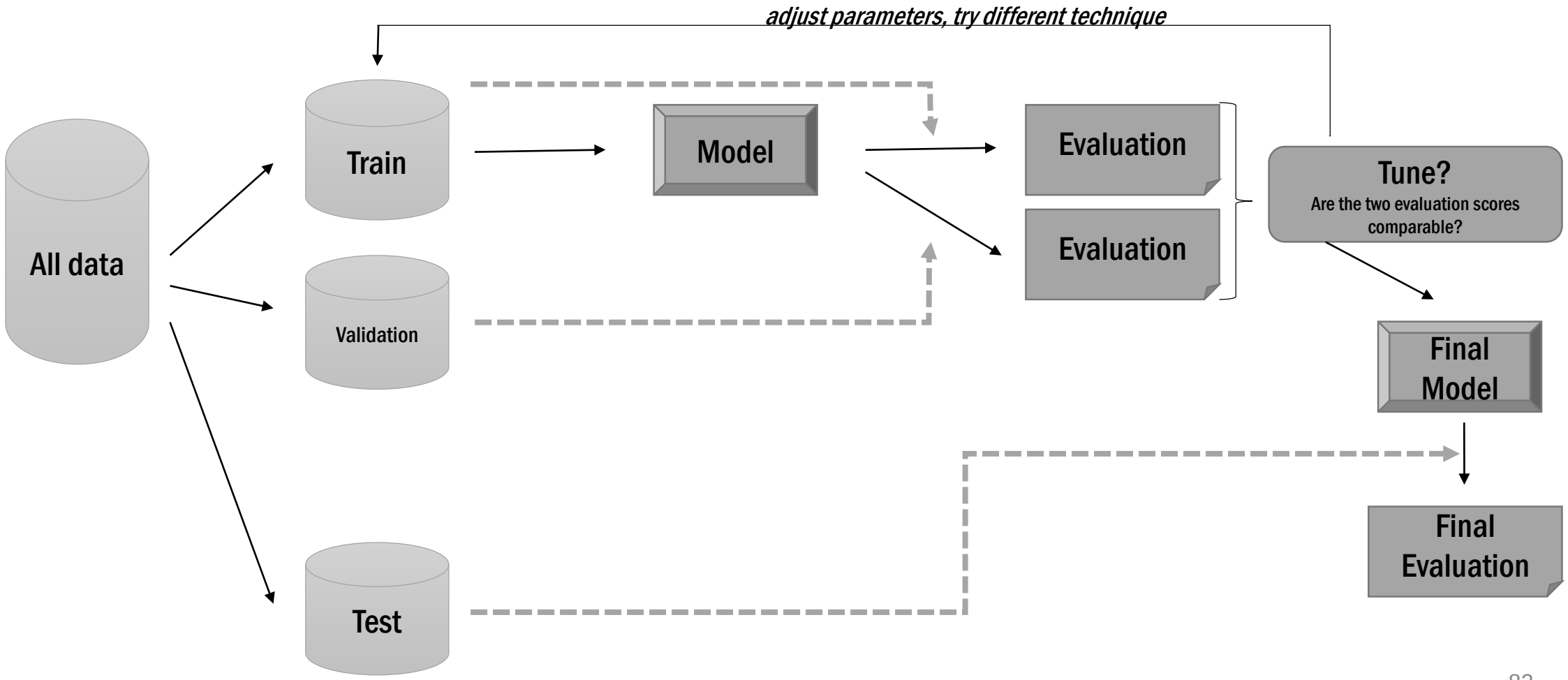
Model validation and evaluation

Utility / profit based?

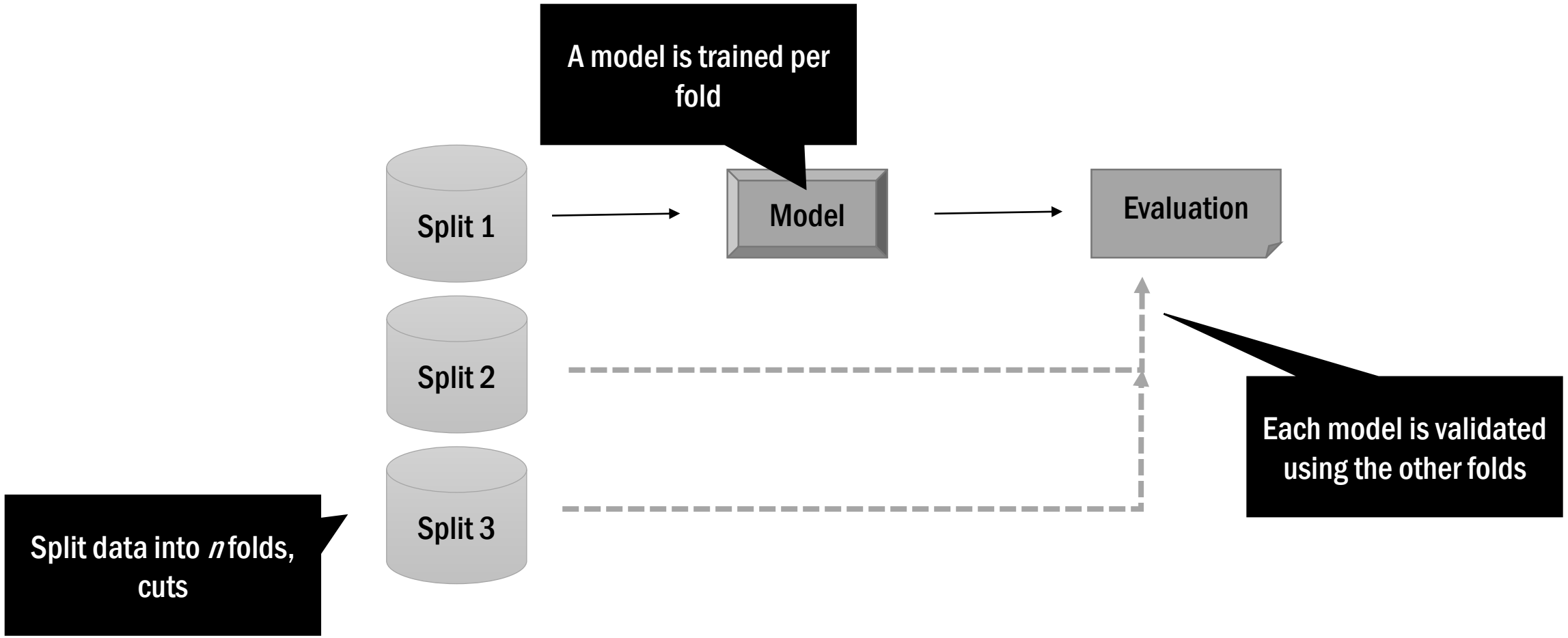




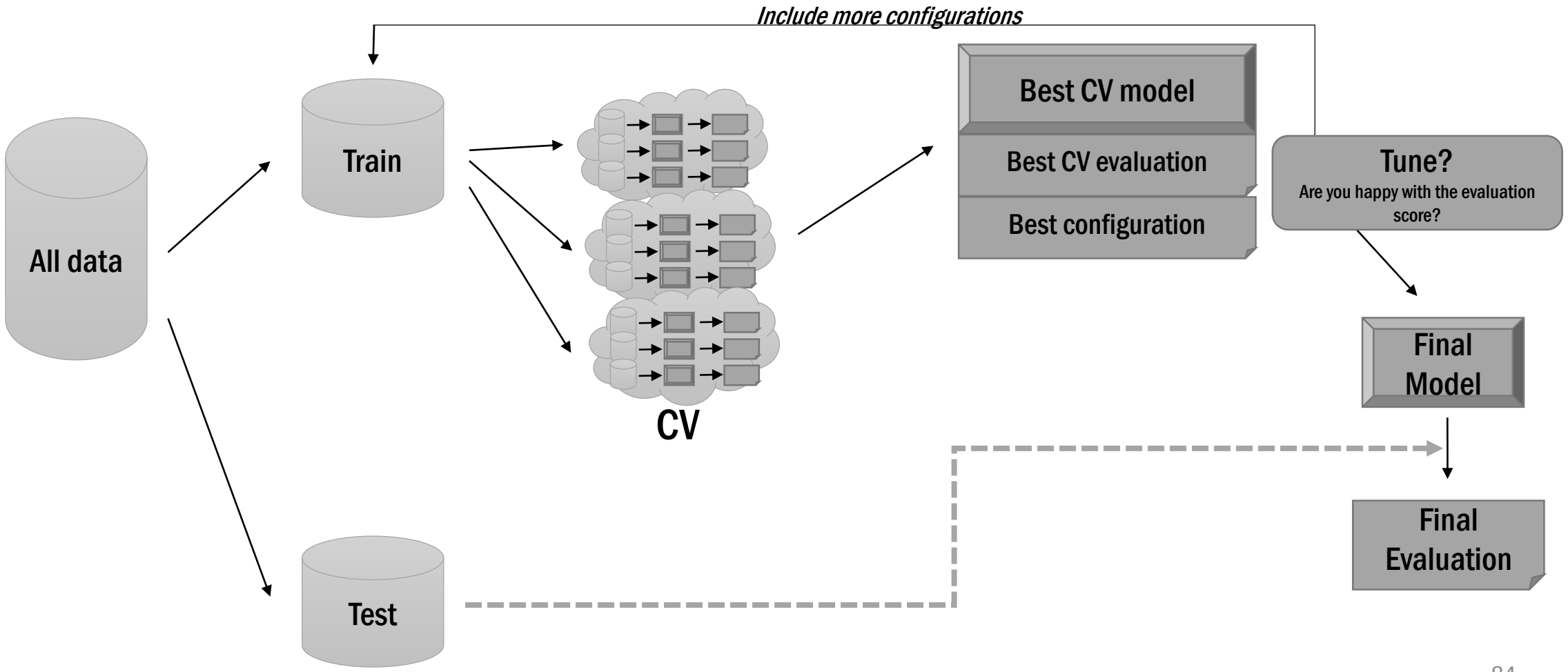
Model validation and evaluation



Model validation and evaluation



Model validation and evaluation



Other concerns

Operational efficiency relates to the effort that is needed to evaluate, monitor, backtest or rebuild the model

- From this perspective, it is quite obvious that a neural network or random forest is less efficient that e.g. a plain vanilla regression model or decision tree
- In some settings like credit card fraud detection, operational efficiency is very important because a decision should be made within a few seconds after the credit card transaction was initiated

Economical cost refers to the cost that is needed to gather the model inputs, run the model and process its outcome(s)

- Also the cost of external data and/or models should be taken into account here
- Calculating the economic return on the analytical model is not a straightforward exercise
- Technical debt and maintenance… will models have to work 1 year from now? 10 years?

Other concerns

Regulatory compliance: e.g. no black-boxes, no use allowed of certain features, premium should be monotonically increasing with sum insured

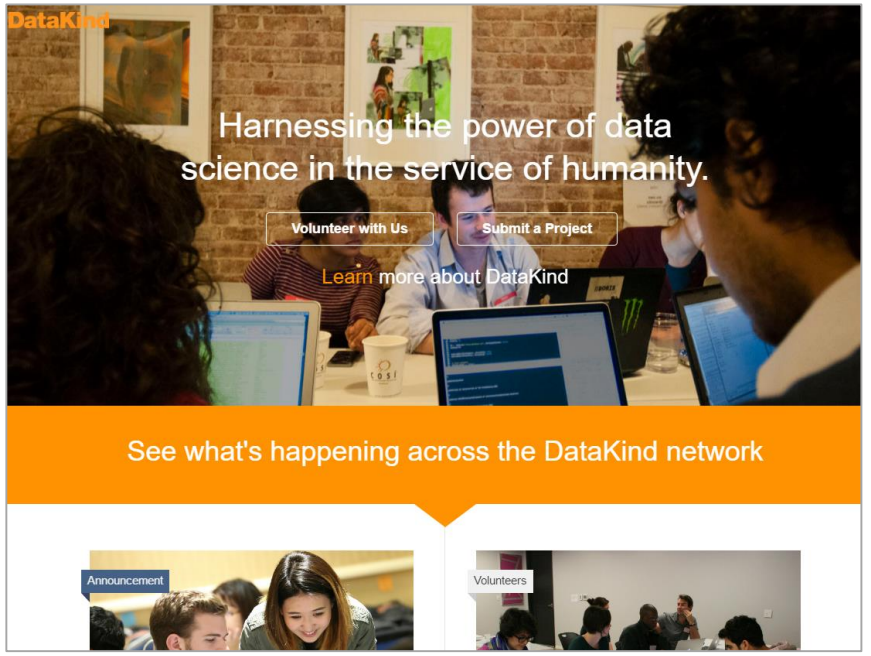
- Questions on ethics…
- Can an algorithm be racist? Sexist?
- “Will Predictive Models Outliers Be The New Socially Excluded?”
- Companies like DataKind, or Bayes Impact
- Concept of “open models”

Data Mining: Where Legality and Ethics Rarely Meet

By Kelly Shermach
Aug 25, 2006 4:00 AM PT

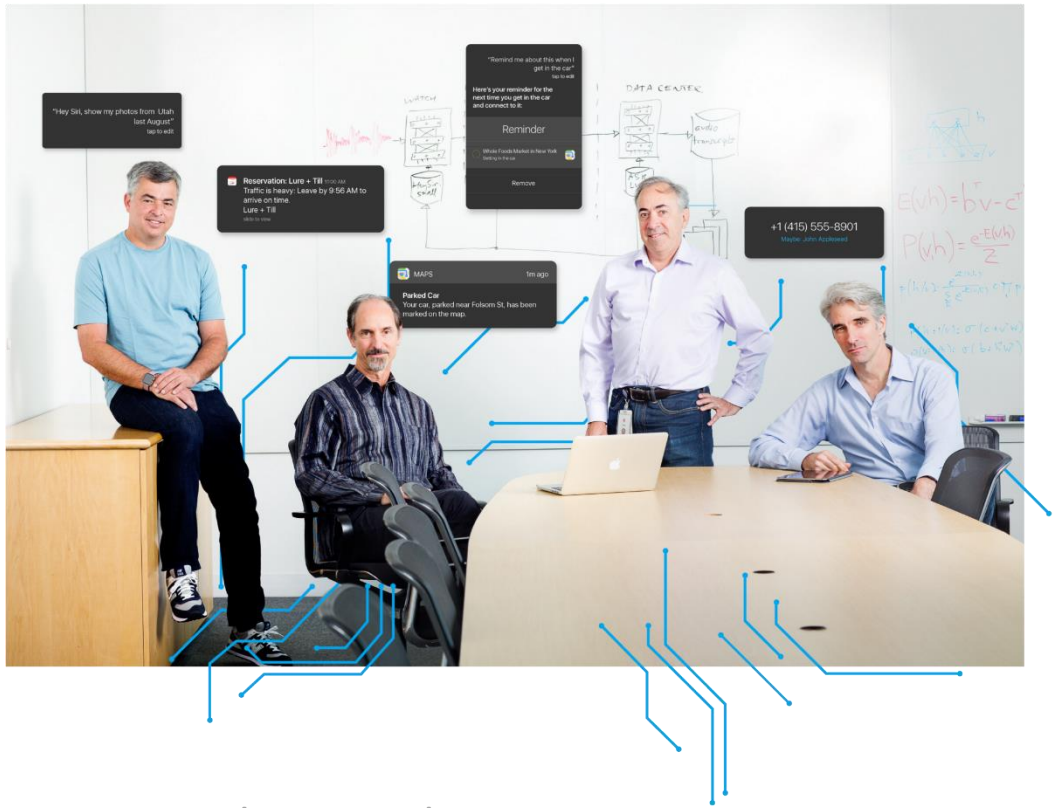
Print *Email*

More than ever, knowingly or unknowingly, consumers disseminate personal data in daily activities. Credit and debit card transactions, ATM visits, Web site browsing and purchases -- even mobile phone use -- all generate data downloaded for analysis and customer profiling. Collectors may use this data to enhance customers' experience, but may also share information with marketers more focused on customer acquisition.



Other concerns

https://backchannel.com/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple-8dbfb131932b#.crky6nt6k



Probably the biggest issue in Apple's adoption of machine learning is how the company can succeed while sticking to its principles on user privacy. The company encrypts user information so that no one, not even Apple's lawyers, can read it (nor can the FBI, even with a warrant). And it boasts about not collecting user information for advertising purposes.

While admirable from a user perspective, Apple's rigor on this issue has not been helpful in luring top AI talent to the company. "Machine learning experts, all they want is data," says a former Apple employee now working for an AI-centric company. "But by its privacy stance, Apple basically puts one hand behind your back. You can argue whether it's the right thing to do or not, but it's given Apple a reputation for not being real hardcore AI folks."

This view is hotly contested by Apple's executives, who say that it's possible to get all the data you need for robust machine learning without keeping profiles of users in the cloud or even storing instances of their behavior to train neural nets. "There has been a false narrative, a false trade-off out there," says Federighi. "It's great that we would be known as uniquely respecting user's privacy. But for the sake of users everywhere, we'd like to show the way for the rest of the industry to get on board here."

> **We have found ways to get that data we need while still maintaining privacy**

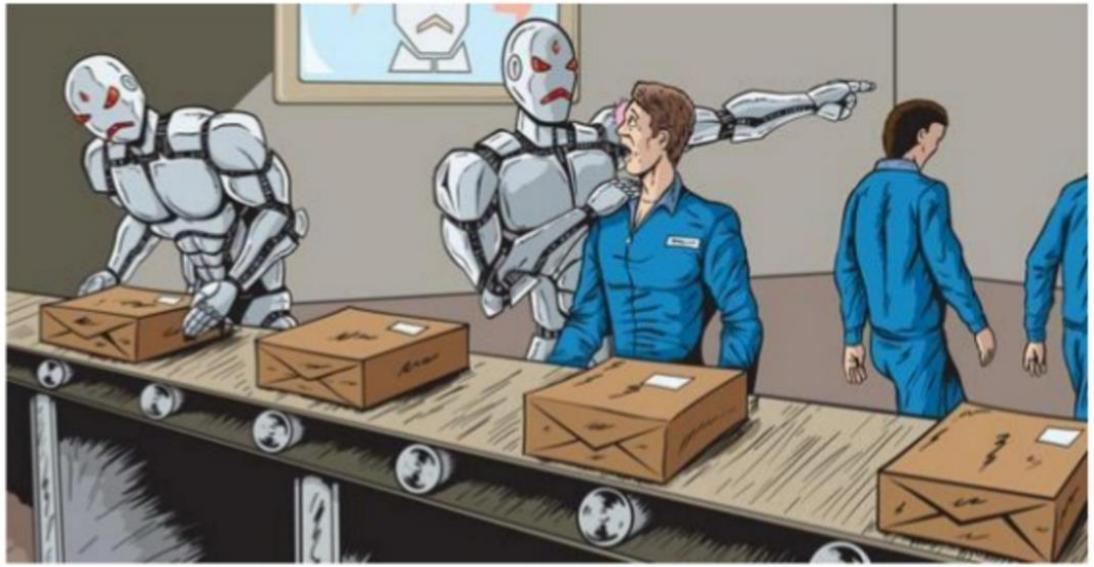
Challenges


INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY
The Economist
How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves
Artificial Intelligence
The promise and the peril
MAY 9TH - 15TH 2015
Worldwide ex UK


Harvard Business Review
You Need an Innovation Strategy
Luxury's Talent Factories
Conquering Digital Distraction
Meet Your New Employee
How to manage the man-machine collaboration
PAGE 57

Challenges

“So why has the robot not replaced the actuary / expert / … yet?”



Challenges

Google failed big data project

All Images News Videos Maps More Search tools

About 3.050.000 results (0,53 seconds)

Hotels.com CTO on why big data projects fail | Data | Comput…
www.computerworlduk.com › Features › Data
Jun 29, 2015 - **Big data projects** have a lot of promise, but the majority **fail**. A recent study found that just 11 percent of corporate leaders in the UK haven't ...

[PDF] Three promises and perils of Big Data - Bain & Company
www.bain.com/.../BAIN_BRIEF_Three_promises_and_perils_of_**Big_D**...
2017, 60% of **Big Data projects** will **fail** to go beyond piloting and experimentation and will be abandoned." Why is history repeating itself? It's not for lack of inter-.

10/26 - Evaluating Big Data Projects - Success and Failure U…
integralleadershipreview.com/10945-evaluating-**big-data-projects**-succes...
With the increasing focus on information analytics and "**big data**," the risks of lagging or **failing projects** are rising due to the complexity of the initiatives and the ...

Why the Majority of Big Data Projects Fail - Qubole
https://www.qubole.com/blog/**big-data**/**big-data-projects-fail**/
Jun 18, 2015 - To truly experience growth in the future, most businesses are turning to **big data**. In many cases, **big data** is seen as the new trend guaranteed ...

The Data Economy: Why do so many analytics projects fail?
analytics-magazine.org/the-**data**-economy-why-do-so-many-analytics-**pr**...
Key considerations for deep analytics on **big data**, learning and insights. ... Accord Gartner, more than half of all analytics **projects fail** because they aren't ...

Where Big Data Projects Fail - Data Science Central
www.**data**sciencecentral.com/profiles/blogs/where-**big-data-projects-fail**
May 11, 2015 - In fact, I predict that half of all **big data projects** will **fail** to deliver against their expectations. **Failure** can happen for many reasons, however ...

Study reveals that most companies are failing at big data | CIO
www.cio.com/.../**big-data**/study-reveals-that-most-companies-are-**failing**-at-**big-data**.ht...
Nov 10, 2015 - Research from PwC and Iron Mountain reports some surprising statistics about how companies are using the **data** they collect.

Where Big Data Projects Fail - Forbes
www.forbes.com/sites/bernardmarr/2015/03/17/where-**big-data-projects-fail**/
Mar 17, 2015 - Over the past 6 months I have seen the number of **big data projects** go up significantly and most of the companies I work with are planning to ...

8 Reasons Big Data Projects Fail - InformationWeek
www.informationweek.com/**big-data**/**big-data**.../8...**big-data-projects-fail**/a/.../129784...
Aug 7, 2014 - Most companies remain on the **big data** sidelines too long, then **fail**. An iterative, start-small approach can help you avoid common pitfalls.

…r Reasons Why Big Data Analytics Projects Fail | Big Data Page by ...
…bage.com/four-reasons-why-**big-data**-analytics-**projects-fail**-or-do-they/
…015 - Surveys say that many **big data** analytics **projects** end in **failure**. Four reasons why, …tion: Did they really **fail**?

…revent Big Data Analytics Failures - Smarter With Gartner
…com/smarterwithgartner/how-to-prevent-**big-data**-analytics-**failures**/
…15 - **Big data** analytics **projects** don't **fail** for a single reason, nor due to technology alone.

…a data-driven culture, big data projects will fail - SearchCIO
…echtarget.com/.../Without-a-**data**-driven-culture-**big-data-projects**-will-**fail**
…a data-driven culture as vital to the success of any **big data project**.

Challenges

- **Mapping the business question** to a technique / setup (there is no one-size fits all)
- Realizing the amount of **effort required in pre-processing**
- **Low amount of training data**, either instances or features
  - Too many features…
  - Huge data imbalance
- **Quality of data**, noise
- Predicting the future is hard (who'd have thought!) – **hard to extrapolate towards the future** for many models (machines are naïve and lazy)
- **Incorporating domain knowledge**, explaining models
- **Strong validation** / backtesting setup requires time and enough data
- **Management:** deployment, monitoring, maintenance, governance

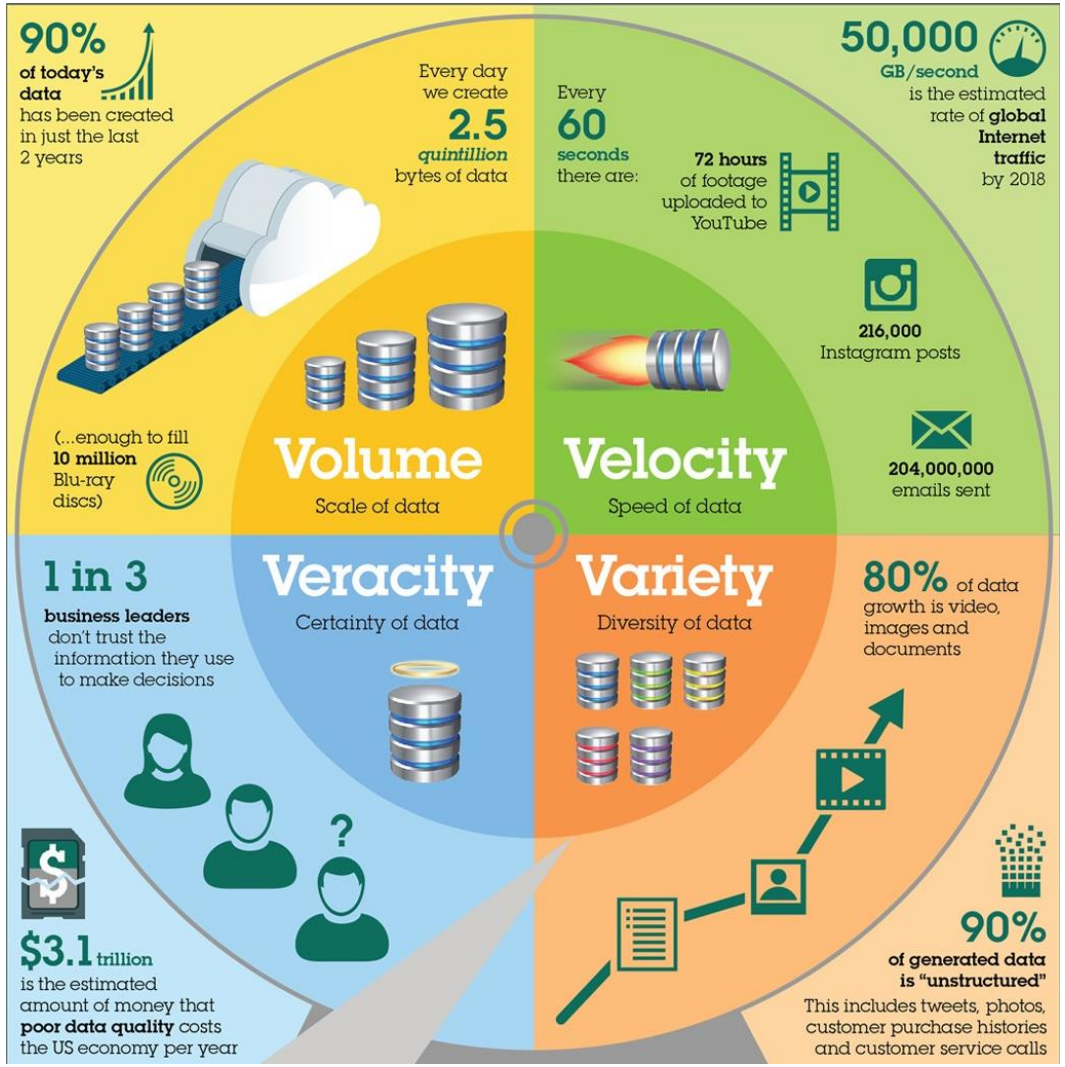
Challenges


INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY
The Economist
How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves
Artificial Intelligence
The promise and the peril
MAY 9TH - 15TH 2015
Worldwide ex UK


Harvard Business Review
You Need an Innovation Strategy
Luxury's Talent Factories
Conquering Digital Distraction
Meet Your New Employee
How to manage the man-machine collaboration

So what about big data?

http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data



So what about big data?

http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data